On the usefulness of Meyer wavelets for deconvolution and density estimation

Jérémie Bigot

March 2009

Abstract

The aim of this paper is to show the usefulness of Meyer wavelets for the classical problem of density estimation and for density deconvolution from noisy observations. By using such wavelets, the computation of the empirical wavelet coefficients relies on the fast Fourier transform and on the fact that Meyer wavelets are band-limited functions. This makes such estimators very simple to compute and this avoids the problem of evaluating wavelets at non-dyadic points which is the main drawback of classical wavelet-based density estimators. Our approach is based on term-by-term thresholding of the empirical wavelet coefficients with random thresholds depending on an estimation of the variance of each coefficient. Such estimators are shown to achieve the same performances of an oracle estimator up to a logarithmic term. These estimators also achieve near-minimax rates of convergence over a large class of Besov spaces. A simulation study is proposed to show the good finite sample performances of the estimator for both problems of direct density estimation and density deconvolution.

Keywords: Density estimation, Deconvolution, Inverse problem, Wavelet thresholding, Random thresholds, Oracle inequalities, Adaptive estimation, Besov space, Minimax rates of convergence.
AMS classifications: Primary 62G07; secondary 42C40, 41A29

Affiliations

Institut de Mathématiques de Toulouse, Université de Toulouse et CNRS (UMR 5219), Jeremie.Bigot@math.univ-toulouse.fr

Acknowledgements

We gratefully acknowledge Yves Rozenholc for providing the Matlab code to compute the model selection estimator.

1 Introduction

Density estimation is a well-known problem in statistics that has been thoroughly studied. It consists in estimating an unknown probability density function f from an independent and identically distributed (iid) sample of random variables X_i for $i = 1, \ldots, n$, with n representing the sample size. Wavelet decomposition is known to be a powerful method for nonparametric estimation, see e.g. Donoho, Johnstone, Kerkyacharian and Picard (1995). The advantages of wavelet methods is their ability in estimating spatially inhomogeneous functions. They can be used to estimate functions in Besov spaces with optimal rates of convergence, and have therefore received special attention in the literature over the last two decades, in particular for density estimation, see e.g. Donoho, Johnstone, Kerkyacharian and Picard (1996) and Vidakovic (1999) for a detailed review on the subject.

For a given scaling function ϕ and mother wavelet ψ, the scaling and wavelet coefficients are usually estimated as Donoho *et al* (1996)

$$\hat{c}_{j_0,k} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j_0,k}(X_i) \text{ and } \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(X_i) \tag{1.1}$$

where $\phi_{j_0,k}(x) = \phi(2^{j_0}x - k)$, $\psi_{j,k}(x) = \psi(2^j x - k)$, and j_0 denotes the usual coarse level of resolution. A hard-thresholding estimator of f then takes the form

$$\hat{f}_n(x) = \sum_k \hat{c}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1}\sum_k \hat{\beta}_{j,k}\mathbb{1}_{\{|\hat{\beta}_{j,k}|\geqslant\tau_{j,k}\}}\psi_{j,k}(x), \tag{1.2}$$

where j_1 is an appropriate frequency cut-off, and where the $\tau_{j,k}$'s are appropriate thresholds (positive numbers) that are possibly level-dependent. However, in practice the computation of the coefficients (1.1) requires some numerical approximation by interpolation as one typically only knows the values of the functions $\phi_{j_0,k}$ and $\psi_{j,k}$ at dyadic points. Various approaches have been used to approximate numerically the coefficients in (1.1). For instance Koo and Kooperberg (2000), Antoniadis, Grégoire and Nason (1999) use a binning method followed by the the standard discrete wavelet transform, while the algorithm proposed in Herrick, Nason and Silverman (2001) is based on an approximation of the scaling coefficients at a sufficiently small level of resolution.

In this paper, we propose to avoid the use of such numerical approximation schemes. This is achieved by using Meyer wavelets which are band-limited functions. Indeed, using such wavelets and thanks to the Plancherel's identity, the empirical coefficients can be easily computed from the Fourier transform of the data $X_i, i = 1, \ldots, n$. Such an approach therefore takes full advantages of the fast Fourier transform and of existing fast algorithms for Meyer wavelet decomposition developed by Kolaczyk (1994).

As Meyer wavelets are band-limited functions, they have been recently used for deconvolution problems in nonparametric regression by Johnstone, Kerkyacharian, Picard and Raimondo (2004), Pensky and Sapatinas (2008), and for density deconvolution by Pensky and Vidakovic (1999), Fan and Koo (2002), Bigot and Van Bellegem (2009). Moreover, the use of such wavelets leads to fast algorithms, see Kolaczyk (1994) and Raimondo and Stewart (2007).

Density deconvolution is the problem of estimating the function f when the observation of the random variable X is contaminated by an independent additive noise. In this case, the observations at hand are a sample of variables $Y_i, i = 1, \ldots, n$ such that

$$Y_i = X_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.3}$$

where X_i are iid variables with unknown density f, and ϵ_i are iid variables with *known* density h which represents some additive noise independent of the X_i's. Density estimation from a noisy sample is of fundamental importance in many practical situations, and applications can be found in communication theory Masry (2003), experimental physics (e.g. Kosarev *et al*, 2003) or econometrics Postel-Vinay and Robin (2002). In this setting, the density of the observed variables Y_i is the convolution of the density f with the density h of the additive noise. Hence, the problem of estimating f relates to nonparametric methods of deconvolution which is a widely studied inverse problem in statistics and signal processing. However the indirect observation of the data leads to different optimality properties, for instance in terms of rate of convergence, than the direct problem of density estimation without an additive error. Standard techniques recently studied for density deconvolution include model selection Comte, Rozenholc and Taupin (2006b), kernel smoothing Carroll and Hall (1988), spline deconvolution Koo (1999), spectral cut-off Johannes (2008) and wavelet thresholding Pensky and Vidakovic (1999), Fan and

Koo (2002), Bigot and Van Bellegem (2009), to name but a few. Again, Meyer wavelets can be used to easily compute estimators of the wavelet coefficients of f by using the Fourier coefficients of the noise density h without using any numerical approximation scheme.

The second contribution of this paper is the use of random thresholds $\tau_{j,k}$. Classically, the thresholds used in wavelet density estimation are deterministic, but such thresholds may be too large in practice as they do not take into account the fact that the variance of each empirical wavelet coefficient $\hat{\beta}_{j,k}$ depends on its location (j,k). The use of random thresholds has been recently proposed in Juditsky and Lambert-Lacroix (2004) in the context of density estimation, and in Reynaud-Bouret and Rivoirard (2008) for Poisson intensity estimation. However, the estimation procedure in Juditsky and Lambert-Lacroix (2004) and Reynaud-Bouret and Rivoirard (2008) is different from the one we propose, since it is based on biorthogonal bases and on the use of the Haar basis to compute the wavelet coefficients. The use of data based threshold exploiting the variance structure of the empirical wavelet coefficients is also proposed in Herrick *et al* (2001), but the theoretical properties of the resulting algorithms are not studied. In this paper, we show that using Meyer wavelets allows one to compute easily an estimation of an upper bound of the variance of the $\hat{\beta}_{j,k}$'s that is then used to compute random thresholds $\tau_{j,k}$.

Then, a third contribution of this paper is that the resulting hard-thresholding estimators are shown to attain the same performances (up to logarithmic terms) of an ideal estimator, called oracle as its computation depends on unknown quantities such as the variance of the $\hat{\beta}_{j,k}$'s or the magnitude of the true wavelet coefficients. Oracle inequalities is an active research area in nonparametric statistics (see e.g. Johnstone (2002), Candès (2005) for detailed expositions) which has recently gained popularity. Deriving an oracle inequality is the problem of bounding the risk of a statistical procedure by the performances of an ideal estimator which represents the best model for the function to recover. Oracle inequalities are currently used in many different contexts in statistics. They have been introduced in Donoho and Johnstone (1994) for nonparametric regression with wavelets, then used by Cavalier, Golubev, Picard and Tsybakov (2002) for inverse problems in the white noise model, by Rigollet (2006), Castellan (2003), Efromovich (2008), Bunea, Tsybakov and Wegkamp (2007) for density estimation problems, and by Reynaud-Bouret and Rivoirard (2008) for estimating the intensity of a Poisson process, to name but a few.

The rest of this paper is organized as follows. In Section 2, we provide some background on Meyer wavelets and we define the corresponding wavelet estimators with random thresholds for both direct density estimation and density deconvolution. In Section 3, it is explained how the performances of such estimates can be compared to those of an oracle estimate, which leads to non-asymptotic oracle inequalities. Asymptotic properties of the estimators are then studied in Section 4. Depending on the problem at hand., these estimators are shown to achieve near-minimax rates of convergence over a large class of Besov spaces. Finally, a simulation study is proposed in Section 5 to evaluate the numerical performances of our estimators and to compare them with other procedures existing in the literature. The proofs of the main theorems are gathered in a technical Appendix.

2 Density estimation with Meyer wavelets

In what follows, it will be assumed that the density function f of the X_i's has a compact support included in $[0,1]$. Of course, assuming that the support of f is included in $[0,1]$ would not hold in many practical applications and this is mainly made for mathematical convenience to simplify the presentation of the estimator. If the range of the data is outside $[0,1]$, one can simply rescale and center them such that they fall into $[0,1]$, and then apply the inverse transformation to the estimated density.

2.1 Wavelet decomposition and the periodized Meyer wavelet basis

Let (ϕ^*, ψ^*) be the Meyer scaling and wavelet function respectively (see Meyer (1992) for further details). It is constructed from a scaling function ϕ^* with Fourier transform

$$\tilde{\phi}(\omega) = \int_{\mathbb{R}} \phi^*(x) e^{-i\omega x} dx = \begin{cases} \frac{\tilde{h}(\omega/2)}{\sqrt{2}} & \text{if} \quad |\omega| \leqslant 4\pi/3, \\ 0 & \text{if} \quad |\omega| > 4\pi/3, \end{cases}$$

where $\tilde{h} : \mathbb{C} \to \mathbb{R}$ is a smooth function chosen as a polynomial of degree 3 in our simulations. Scaling and wavelet function at scale $j \geq 0$ are defined by

$$\phi^*_{j,k}(x) = 2^{j/2}\phi^*(2^j x - k) \text{ and } \psi^*_{j,k}(x) = 2^{j/2}\psi^*(2^j x - k),\ k = 0, \ldots, 2^j - 1.$$

As in Johnstone *et al* (2004) , one can then define the periodized Meyer wavelet basis of $L^2([0,1])$ (the space of squared integrable functions on $[0,1]$) by periodizing the functions (ϕ^*, ψ^*) i.e.

$$\phi_{j,k}(x) = 2^{j/2} \sum_{i \in \mathbb{Z}} \phi^*(2^j(x+i) - k) \text{ and } \psi_{j,k}(x) = 2^{j/2} \sum_{i \in \mathbb{Z}} \psi^*(2^j(x+i) - k),\ k = 0, \ldots, 2^j - 1.$$

For any function f of $L^2([0,1])$, its wavelet decomposition can be written as:

$$f = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

where $c_{j_0,k} = \langle f, \phi_{j_0,k} \rangle = \int_0^1 f(u)\phi_{j_0,k}(u)du$, $\beta_{j,k} = \langle f, \psi_{j,k} \rangle = \int_0^1 f(u), \psi_{j,k}(u)du$ and $j_0 \geq 0$ denotes the usual coarse level of resolution. Moreover, the L^2 norm of f is given by

$$\|f\|^2 = \sum_{k=0}^{2^{j_0}-1} c^2_{j_0,k} + \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \beta^2_{j,k}.$$

Meyer wavelets can be used to efficiently compute the coefficients $c_{j,k}$ and $\beta_{j,k}$ by using the Fourier transform. Indeed, let $e_\ell(x) = e^{2\pi i \ell x}, \ell \in \mathbb{Z}$ and denote by $f_\ell = \langle f, e_\ell \rangle = \int_0^1 f(u)e^{-2\pi i \ell u}du$ the Fourier coefficients of f supposed to be a function in $L^2([0,1])$. By the Plancherel's identity, we obtain that

$$\beta_{j,k} = \langle f, \psi_{j,k} \rangle = \sum_{\ell \in C_j} \psi^{j,k}_\ell f_\ell, \tag{2.1}$$

where $\psi^{j,k}_\ell = \langle \psi_{j,k}, e_\ell \rangle$ denote the Fourier coefficients of $\psi_{j,k}$ and $C_j = \{\ell \in \mathbb{Z}; \psi^{j,k}_\ell \neq 0\}$. As Meyer wavelets $\psi_{j,k}$ are band-limited C_j is a finite subset set of $[-2^{j+2}c_0, -2^j c_0] \cup [2^j c_0, 2^{j+2}c_0]$ with $c_0 = 2\pi/3$ (see Johnstone *et al* (2004)).

2.2 The case of direct density estimation

Based on a sample $X_1, \ldots, X_n$, an unbiased estimator of f_ℓ is given by $\frac{1}{n}\sum_{m=1}^n \exp(-2\pi i \ell X_m)$ which yields an unbiased estimator of $\beta_{j,k}$ given by

$$\hat{\beta}_{j,k} = \sum_{\ell \in C_j} \psi^{j,k}_\ell \left(\frac{1}{n} \sum_{m=1}^n e^{-2\pi i \ell X_m} \right). \tag{2.2}$$

The coefficients (2.2) can therefore be easily calculated by combining the fast Fourier transformation of the data with the fast algorithm of Kolaczyk (1994) which relies on the fact that the first sum in (2.2)

only involves a finite number of terms. Equation (2.2) also shows that using Meyer wavelets, which are band-limited functions, avoids the use of numerical schemes to approximate the computation of the coefficients (1.1) as one typically only knows the values of the functions $\phi_{j_0,k}$ and $\psi_{j,k}$ at dyadic points. We define the estimators of the scaling coefficients $c_{j_0,k}$ analogously, with ϕ instead of ψ and $C_{j_0} = \{\ell \in \mathbb{Z}; \phi_\ell^{j_0,k} \neq 0\}$ instead of C_j.

2.3 The case of density deconvolution

Consider now the problem (1.3) of density deconvolution. Denote by $f_\ell = \langle f, e_\ell \rangle$ the Fourier coefficients of f, and by $h_\ell = \int_{\mathbb{R}} h(u) e_\ell(u) du$ the Fourier coefficients of the error density h. Since $f_\ell = \mathbb{E}(e^{-2\pi i \ell X_1})$ and $h_\ell = \mathbb{E}(e^{-2\pi i \ell \epsilon_1})$, it follows by independence that $\mathbb{E}(e^{-2\pi i \ell Y_1}) = f_\ell h_\ell$. This equality and the Plancherel's identity (2.1) therefore implies that an unbiased estimator of $\beta_{j,k}$ is given by (assuming that $h_\ell \neq 0$ for all $\ell \in \mathbb{Z}$)

$$\hat{\beta}_{j,k} = \sum_{\ell \in C_j} \tilde{\psi}_\ell^{j,k} \left(\frac{1}{n} \sum_{m=1}^{n} e^{-2\pi i \ell Y_m} \right) \text{ where } \tilde{\psi}_\ell^{j,k} = \frac{\psi_\ell^{j,k}}{h_\ell}. \tag{2.3}$$

It is well-known that the difficulty of the deconvolution problem is quantified by the smoothness of the error density h. The so-called ill-posedness of such inverse problems depends on how fast the Fourier coefficients h_ℓ tend to zero. Depending on the decay of these coefficients, the estimation of f_ℓ will be more or less accurate. In this paper, we consider the case where the h_ℓ's have a polynomial decay which is usually referred to as ordinary smooth convolution (see e.g. Fan (1991)):

Assumption 2.1 *The Fourier coefficients of h have a polynomial decay which means that there exists a real $\nu \geqslant 0$ and two positive constants $C_{\min}, C_{\max}$ such that for all $\ell \in \mathbb{Z}$, $C_{\min}|\ell|^{-\nu} \leq |h_\ell| \leq C_{\max}|\ell|^{-\nu}$.*

The rates of convergence that can be expected from a wavelet estimator depend on such smoothness assumptions and are well-studied in the literature, and we refer to Pensky and Vidakovic (1999), Fan and Koo (2002) for further details.

Finally note that to simplify the presentation, we prefer to use the same notation $\hat{\beta}_{j,k}$ and $\hat{c}_{j_0,k}$ for both problems of direct density estimation and density deconvolution.

2.4 Thresholding of the empirical wavelet coefficients

Based on an estimation of the scaling and wavelet coefficients, a linear wavelet estimator of f is of the form $\hat{f}_L = \sum_{k=0}^{2^{j_0}-1} \hat{c}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k}$. For an appropriate choice of j_1 one can show that $\hat{f}_L$ achieves optimal rates of convergence among the class of linear estimators. Typically, if f belongs to a Sobolev space H^s with smoothness order s, then for direct density estimation the choice $2^{j_1} \approx n^{\frac{1}{2s+1}}$ yields optimal rates of convergence for the quadratic risk, see Donoho *et al* (1996), Juditsky and Lambert-Lacroix (2004) for further details . However, this choice is not adaptive because it depends on the unknown smoothness s of f. It is well known that adaptivity can be obtained by using nonlinear estimators based on appropriate thresholding of the estimated wavelet coefficients. A non-linear estimator by hard-thresholding is defined by

$$\hat{f}_n = \sum_{k=0}^{2^{j_0}-1} \hat{c}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k} \mathbb{1}_{\{|\hat{\beta}_{j,k}| \geqslant \tau_{j,k}\}} \psi_{j,k} \tag{2.4}$$

where the $\tau_{j,k}$'s are appropriate thresholds (positive numbers). Various choices for j_1 and the threshold $\tau_{j,k}$ have been proposed. In the case of direct density estimation, Donoho *et al* (1996) recommended to

take level-dependent threshold $\tau_{j,k} \sim \sqrt{j/n}$ and $2^{j_1} \sim \frac{n}{\log(n)}$. For density deconvolution, one possible calibration in ordinary smooth deconvolution is $2^{j_1} \sim n^{\frac{1}{2\nu+1}}$ and $\tau_{j,k} \sim \frac{2^{\nu j}}{\sqrt{n}}$, see Pensky and Vidakovic (1999). The choices $\tau_{j,k} \sim 2^{\nu j}\sqrt{j/n}$ and $\tau_{j,k} \sim 2^{\nu j}\sqrt{2\frac{\log(n)}{n}}$ have also been considered in Fan and Koo (2002) and Bigot and Van Bellegem (2009) respectively.

However, such thresholds may be too large in practice as they do not take into account the fact that the variance of each empirical wavelet coefficient $\hat{\beta}_{j,k}$ depends on its location (j,k). Consider the problem of density deconvolution and let us denote the variance of $\hat{\beta}_{j,k}$ by

$$\sigma_{j,k}^2 = \mathbb{E}(\hat{\beta}_{j,k} - \beta_{j,k})^2.$$

Let us also denote by $\tilde{\psi}_{j,k}$ the function defined for $y \in \mathbb{R}$ by

$$\tilde{\psi}_{j,k}(y) = \sum_{\ell \in C_j} \tilde{\psi}_\ell^{j,k} e^{-2\pi i \ell y}.$$

By definition, it follows that $\hat{\beta}_{j,k} = \frac{1}{n}\sum_{m=1}^n \tilde{\psi}_{j,k}(Y_m)$ and thus $\sigma_{j,k}^2 = \frac{1}{n}\mathrm{Var}(\tilde{\psi}_{j,k}(Y_1))$. Hence, a simple upper bound for $\sigma_{j,k}^2$ is $V_{jk} = \frac{1}{n}\mathbb{E}(\tilde{\psi}_{j,k}(Y_1)^2)$. Then, simple algebra shows that in the case of density deconvolution

$$V_{j,k} = \frac{1}{n}\int_{\mathbb{R}} |\tilde{\psi}_{j,k}(y)|^2 f^Y(y)dy = \frac{1}{n}\sum_{\ell,\ell' \in C_j} \tilde{\psi}_\ell^{j,k}\overline{\tilde{\psi}_{\ell'}^{j,k}} f_{\ell-\ell'}^Y, \tag{2.5}$$

with $f^Y(y) = \int_0^1 f(u)h(y-u)du$ and $f_{\ell-\ell'}^Y = \mathbb{E}e^{-2\pi i(\ell-\ell')Y_1}$. An unbiased estimator of $V_{j,k}$ is thus given by

$$\hat{V}_{j,k} = \frac{1}{n}\sum_{\ell,\ell' \in C_j} \tilde{\psi}_\ell^{j,k}\overline{\tilde{\psi}_{\ell'}^{j,k}} \left(\frac{1}{n}\sum_{m=1}^n e^{-2\pi i(\ell-\ell')Y_m}\right) = \frac{1}{n^2}\sum_{m=1}^n \left|\sum_{\ell \in C_j} \tilde{\psi}_\ell^{j,k} e^{-2\pi i \ell Y_m}\right|^2. \tag{2.6}$$

Similar computations can be made for the case of direct density estimation with $\psi_\ell^{j,k}$ instead $\tilde{\psi}_\ell^{j,k}$, $f_{\ell-\ell'}$ instead of $f_{\ell-\ell'}^Y$, and X_m instead of Y_m in the above equations (2.5) and (2.6). Note that $\hat{V}_{j,k}$ can also be written as $\hat{V}_{j,k} = \frac{1}{n^2}\sum_{m=1}^n \left|\tilde{\psi}_{j,k}(Y_m)\right|^2$, but its calculation is obtained from the Fourier coefficients $(\tilde{\psi}_\ell^{j,k})_{\ell \in C_j}$ and not from $\tilde{\psi}_{j,k}$ whose computation at non dyadic points requires numerical approximation. Alternatively, one could also use an estimation of the variance $\sigma_{j,k}^2$ given by

$$\hat{\sigma}_{j,k}^2 = \hat{V}_{j,k} - \frac{1}{n}|\hat{\beta}_{j,k}|^2,$$

instead of the upper bound $V_{j,k}$. However, this does not change significantly our results since $\hat{\sigma}_{j,k}^2$ and $\hat{V}_{j,k}$ are very close for n sufficiently large. Moreover, oracle inequalities are simpler to derive using an estimated upper bound for $\sigma_{j,k}^2$.

A thresholding rule is usually chosen by controlling the probability of deviation of $\hat{\beta}_{j,k}$ from the true wavelet coefficient $\beta_{j,k}$. From Lemma 5.3 (see the Appendix) one has that for any positive x, $\mathbb{P}\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \sqrt{2V_{jk}x} + \frac{\eta_j}{3n}x\right) \leq 2\exp(-x)$ where

$$\eta_j = \sum_{\ell \in C_j} |\tilde{\psi}_\ell^{j,k}|, \tag{2.7}$$

which would suggest to take a threshold of the form

$$\tau^*_{j,k} = \sqrt{2\delta \log(n) V_{j,k}} + \frac{\delta \log(n)}{3n}\eta_j,$$

where $\delta > 0$ is a tuning parameter. Thinking of the classical universal threshold, one would take $\delta = 1$. However, this choice can be too conservative and the results of this paper shows that it is possible to take δ smaller that 1. Moreover, it is shown that the choice of this tuning parameter depends on the highest resolution level j_1 and the degree of ill-posedness ν in the case of density deconvolution. Throughout the paper, we discuss the choice of δ and finally propose data-based values for its calibration.

Obviously $\tau^*_{j,k}$ is an ideal threshold as $V_{j,k}$ is unknown. Based on Lemma 5.4 (see the Appendix) which gives a control on the probability of deviation of $\hat{V}_{j,k}$ from $V_{j,k}$, we propose to use the following random thresholds

$$\tau_{j,k} = \sqrt{2\delta \log(n) \left(\hat{V}_{j,k} + \sqrt{2\delta \log(n) \hat{V}_{j,k} \frac{\eta_j^2}{n^2}} + \delta \log(n) \kappa \frac{\eta_j^2}{n^2} \right)} + \frac{\delta \log(n)}{3n}\eta_j, \tag{2.8}$$

where $\kappa = \frac{4}{3} + \sqrt{\frac{5}{3}}$

Again, for direct density estimation, we take the same thresholds with $\psi_\ell^{j,k}$ instead of $\tilde{\psi}_\ell^{j,k}$ to compute η_j in (2.7). The above choice for $\tau_{j,k}$ resembles to the universal threshold $\sigma\sqrt{2\log(n)}$ proposed by Donoho and Johnstone (1994) in the context of nonparametric regression with homoscedastic variance σ^2. Here we exploit the fact that in the context of density estimation the variance of a wavelet coefficient depends on its location (j,k) and has to be estimated.

The additive terms $\frac{\delta \log(n)}{3n}\eta_j$ and $\sqrt{2\delta \log(n) \hat{V}_{j,k} \frac{\eta_j^2}{n^2}} + \delta \log(n) \kappa \frac{\eta_j^2}{n^2}$ will allow us to derive oracle inequalities. Indeed, in Section 3, we compare the quadratic risk of $\hat{f}_n$ to the risk of the following oracle estimator

$$\tilde{f}_n = \sum_{k=0}^{2^{j_0}-1} \hat{c}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k} \mathbb{1}_{\{|\beta_{j,k}|^2 \geqslant \sigma^2_{j,k}\}} \psi_{j,k} \tag{2.9}$$

Note that $\tilde{f}_n$ is an ideal estimator that can not be computed in practice as its depends on the unknown coefficients $\beta_{j,k}$ of f and the unknown variance terms σ^2_{jk}. However, we shall use it as a benchmark to assess the quality of our estimator. The quadratic risk of $\tilde{f}_n$ is

$$\mathbb{E}\|\tilde{f}_n - f\|^2 = \sum_{k=0}^{2^{j_0}-1} \sigma^2_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \min(\beta^2_{j,k}, \sigma^2_{j,k}) + \sum_{j=j_1+1}^{+\infty} \sum_{k=0}^{2^j-1} \beta^2_{j,k}, \tag{2.10}$$

where $\sigma^2_{j_0,k} = \mathbb{E}(\hat{c}_{j_0,k} - c_{j_0,k})^2$. Equation (2.10) shows that we retrieve the classical formula for the quadratic risk of an oracle estimator given in Donoho and Johnstone (1994) except that the variance term $\sigma^2_{j,k}$ is not constant as in standard nonparametric regression with homoscedastic variance.

Data-driven thresholds based on an estimation of the variance of the empirical wavelet coefficients have already been proposed in Juditsky and Lambert-Lacroix (2004) in the context of density estimation. However, the estimation procedure in Juditsky and Lambert-Lacroix (2004) is different from ours since it is based on biorthogonal bases and on the use of the Haar basis to compute the wavelet coefficients. Note that our choice for $\tau_{j,k}$ is similar to the random threshold in Juditsky and Lambert-Lacroix (2004), but the addition of the terms $\sqrt{2\delta \log(n) \hat{V}_{j,k} \frac{\eta_j^2}{n^2}} + \delta \log(n) \kappa \frac{\eta_j^2}{n^2}$ and $\frac{\delta \log(n)}{3n}\eta_j$ will allow us to compare the

performances of $\hat{f}_n$ with those of the oracle $\tilde{f}_n$. The addition of similar deterministic and stochastic terms is also proposed in Reynaud-Bouret and Rivoirard (2008) to derive oracle inequalities in the context of Poisson intensity estimation.

3 Oracle inequalities

To derive oracle inequalities, we need a further smoothness assumption on the error density h:

Assumption 3.1 *There exists a constant $C > 0$ and a real $\rho > 1$ such that the density h satisfies $h(x) \leq \frac{C}{1+|x|^\rho}$ for all $x \in \mathbb{R}$.*

Obviously, the above condition for h is not very restrictive as h is by definition an integrable function on $\mathbb{R}$. For a bounded function $f \in L^2([0,1])$ we denote by $\|f\|_\infty = \sup_{x\in[0,1]}\{|f(x)|\}$ its supremum norm. Let us also define the following class of densities

$$D^2([0,1]) = \{f \in L^2([0,1]), \text{ with } f \geq 0 \text{ and } \int_0^1 f(x)dx = 1\}.$$

3.1 The case of direct density estimation

The following theorem states that for appropriate choices of j_1, j_0 and the tuning parameter δ, then the estimator $\hat{f}_n$ behaves essentially as the oracle $\tilde{f}_n$ up to logarithmic terms.

Theorem 3.1 *Assume that $f \in D^2([0,1])$ with $\|f\|_\infty < +\infty$. Let $\alpha \geq 0$ and $1/2 \geq \eta > 0$ be some fixed constants. For any $n \geq \exp(1)$, define $j_1 = j_1(n)$ to be the integer such that $2^{j_1} > n^\eta(\log n)^\alpha \geq 2^{j_1-1}$, and $j_0 = j_0(n)$ to be the integer such that $2^{j_0} > \log(n) \geq 2^{j_0-1}$ and suppose that η and α are such that $j_1 \geq j_0$. Assume that $\delta > \eta$, and take the random thresholds $\tau_{j,k}$ given by equation (2.8). Then, the estimator $\hat{f}_n$ satisfies the following oracle inequality*

$$\mathbb{E}\|\hat{f}_n - f\|^2 \leq C_1(\delta)\left[\sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{j,k}^2, \log(n)\sigma_{j,k}^2) + \sum_{j=j_1+1}^{+\infty}\sum_{k=0}^{2^j-1}\beta_{j,k}^2\right] + C_2(\delta)\Gamma_{n,1}, \quad (3.1)$$

where

$$\Gamma_{n,1} = \frac{\log(n)}{n}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2 + \max(\|f\|_\infty, 1)\max((\log n)^\alpha, 1)\frac{(\log n)^\alpha}{n} + \frac{(\log n)^{2+2\alpha}}{n^{2(1-\eta)}}$$

and $C_1(\delta)$ and $C_2(\delta)$ are two positive constants not depending on n and f, and such that $\lim_{\delta\to\eta} C_1(\delta) = \lim_{\delta\to\eta} C_2(\delta) = +\infty$.

The above inequality (3.1) shows that the performances of $\hat{f}_n$ mimic those of the oracle $\tilde{f}_n$ in term of quadratic risk, see equation (2.10), up to a logarithmic term. The additive term $\Gamma_{n,1}$ depends on two hyperparameters α and η which are used to control the effect of the choice of the highest resolution level j_1 and the tuning parameter δ on the performances of the estimator. Moreover, the above inequality tends to show that the performances of $\hat{f}_n$ deteriorates as δ tends to η. Thinking of the classical universal threshold, one would like to take $\delta = 1$. However, if one sets $\eta = 1/2$ and $\alpha = 0$, the additive term $\Gamma_{n,1}$ is bounded by $\frac{(\log n)^2}{n}$ which is rate typically faster that the decay of the oracle risk (2.10) when f belongs to a Sobolev or a Besov space. Hence, Theorem 3.1 shows that if one chooses $j_1 = \lfloor \eta \log_2(n) \rfloor + 1$ with $\eta \leq 1/2$, then it is possible to take δ smaller than 1. Choosing carefully such hyperparameters is of fundamental importance, and a detailed study is therefore proposed in Section 5 to validate the results

of Theorem 3.1, and to analyze the risk of $\hat{f}_n$ as a function of the resolution level j_1 and the tuning constant δ.

Classically, the level j_1 is chosen as the integer j_1 such that $2^{j_1} \geq \frac{n}{\log(n)} \geq 2^{j_1-1}$ (see e.g. Donoho *et al* (1996)). The results of this paper shows that one can use smaller level 2^{j_1} of the order $n^{\eta}(\log n)^{\alpha}$. For $\eta \leq 1/2$, the price to pay is a slightly lower rate for the additive term $\Gamma_{n,1}$ in the oracle inequality (3.1) which is of the order $\frac{(\log n)^2}{n}$ instead of the rate $\frac{1}{n}$ as classically obtained for the additive term when deriving oracle inequalities. Note that a similar result in the context of Poisson intensity estimation is also given in Reynaud-Bouret and Rivoirard (2008). In particular Reynaud-Bouret and Rivoirard (2008) obtain an additive term of the order $\frac{1}{n}$ but to derive such a result their proof relies heavily on the fact that the wavelet basis they use (the Haar basis) is such that $\inf_{x\in[0,1]} |\psi(x)| > 0$ which is not the case for Meyer wavelets.

3.2 The case of density deconvolution

Consider now the problem of density deconvolution under Assumption 2.1 of ordinary smooth deconvolution.

Theorem 3.2 *Assume that $f \in D^2([0,1])$ with $\|f\|_\infty < +\infty$, and that h satisfies Assumption 2.1 and Assumption 3.1. Let $\alpha \geq 0$ and $1/2 \geq \eta > 0$ be some fixed constants. For any $n > \exp(1)$, define $j_1 = j_1(n)$ to be the integer such that $2^{j_1} > n^{\eta/(\nu+1)}(\log n)^{\alpha} \geq 2^{j_1-1}$, and $j_0 = j_0(n)$ to be the integer such that $2^{j_0} > \log(n) \geq 2^{j_0-1}$, and suppose that η and α are such that $j_1 \geq j_0$. Assume that $\delta > \eta\left(1+\frac{\nu}{\nu+1}\right)$, and take the random thresholds $\tau_{j,k}$ given by equation (2.8). Then, the estimator $\hat{f}_n$ satisfies the following oracle inequality*

$$\mathbb{E}\|\hat{f}_n - f\|^2 \leq C_3(\delta)\left[\sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{j,k}^2, \log(n)\sigma_{j,k}^2) + \sum_{j=j_1+1}^{+\infty}\sum_{k=0}^{2^j-1}\beta_{j,k}^2\right] + C_4(\delta)\Gamma_{n,2}, \quad (3.2)$$

where

$$\Gamma_{n,2} = \frac{\log(n)}{n}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2 + \max(\|f\|_\infty, 1)\max((\log n)^{\alpha}, 1)\frac{(\log n)^{\alpha(2\nu+1)}}{n} + \frac{(\log n)^{2+2\alpha+2\alpha\nu}}{n^{2(1-\eta)}}$$

and $C_3(\delta)$ and $C_4(\delta)$ are two positive constants not depending on n and f, and such that $\lim_{\delta\to\eta\left(1+\frac{\nu}{\nu+1}\right)} C_3(\delta) = \lim_{\delta\to\eta\left(1+\frac{\nu}{\nu+1}\right)} C_4(\delta) = +\infty$.

Hence, the above theorem shows that in the case of density deconvolution then the estimator also behaves as an oracle estimate up to logarithmic terms, and that the performances of the estimator tend to deteriorate as δ tends to $\eta\left(1+\frac{\nu}{\nu+1}\right)$. Similar comments to those given for Theorem 3.1 can be made. If one chooses $\eta = 1/2$, then the additive term $\Gamma_{n,2}$ is of the order $\frac{(\log n)^{2+2\alpha+2\alpha\nu}}{n}$, and δ has to be greater than $\left(1/2+\frac{\nu}{2\nu+2}\right)$. Hence, this again shows that one can take a value for δ smaller than 1. However the choice of δ is typically larger for deconvolution than in the direct case, as it is controlled by the degree ν of ill-posedness.

In deconvolution problems, the high-frequency cut-off j_1 is usually related to the ill-possedness ν of the inverse problem and $2^{j_1} = \mathcal{O}\left((\frac{n}{\log(n)})^{1/(2\nu+1)}\right)$ is a typical choice for various estimators proposed in the literature (see e.g. Johnstone *et al* (2004), Pensky and Sapatinas (2008), Bigot and Van Bellegem (2009)). This is a standard fact that a smaller j_1 should be used for ill-posed inverse problems than in the direct case. Again we have introduced hyperparameters α and η to control the effect of the choice of

the highest resolution level j_1. Understanding the influence of the choice of these hyperparameters on the quality of the estimator is a fundamental issue, and a detailed simulation study is thus proposed in Section 5 to validate the results of Theorem 3.2.

4 Asymptotic properties and near-minimax optimality

It is well known that Besov spaces for periodic functions in $L^2([0,1])$ can be characterized in terms of wavelet coefficients (see e.g. Johnstone *et al* (2004)). Let $s > 0$ denote the usual smoothness parameter, then for the Meyer wavelet basis and for a Besov ball $B^s_{p,q}(A)$ of radius $A > 0$ with $1 \leq p, q \leq \infty$, one has that for $s + 1/2 - 1/p \geq 0$

$$B^s_{p,q}(A) = \left\{ f \in L^2([0,1]) : \left(\sum_{k=0}^{2^{j_0}-1} |c_{j_0,k}|^p \right)^{\frac{1}{p}} + \left(\sum_{j=j_0}^{+\infty} 2^{j(s+1/2-1/p)q} \left(\sum_{k=0}^{2^j-1} |\beta_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \leq A \right\}$$

with the respective above sums replaced by maximum if $p = \infty$ or $q = \infty$.

The condition that $s + 1/2 - 1/p \geq 0$ is imposed to ensure that $B^s_{p,q}(A)$ is a subspace of $L^2([0,1])$, and we shall restrict ourselves to this case in this paper. Besov spaces allow for more local variability in local smoothness than is typical for functions in the usual Hölder or Sobolev spaces. For instance, a real function f on $[0,1]$ that is piecewise continuous, but for which each piece is locally in C^s, can be an element of $B^s_{p,p}(A)$ with $1 \leq p < 2$, despite the possibility of discontinuities at the transition from one piece to the next. Note that if $s \geq 1$ is not an integer, then $B^s_{2,2}(A)$ is equivalent to a Sobolev ball of order s, and that the space $B^s_{p,q}(A)$ with $1 \leq p < 2$ contains piecewise smooth functions with local irregularities such as discontinuities.Finally let us introduce the following space of densities

$$D^s_{p,q}(A) = D^2([0,1]) \cap B^s_{p,q}(A).$$

The following theorems show that for either the problem of direct density estimation or density deconvolution, the estimator $\hat{f}_n$ is asymptotically near-optimal (in the minimax sense) up to a logarithmic factor over a wide range of Besov balls.

To simplify the presentation, all the asymptotic properties of $\hat{f}_n$ are given in the case $\eta = 1/2$ and $\alpha = 0$ where η, α are the hyperparameters introduced in Theorem 3.1 and Theorem 3.2.

4.1 The case of direct density estimation

For $s > 0$ and $1 \leq p \leq \infty$ and let us define

$$p' = \min(2, p) \text{ and } s^* = s + 1/2 - 1/p'.$$

Then the following result holds.

Theorem 4.1 *Assume that the conditions of Theorem 3.1 hold with $\eta = 1/2$ and $\alpha = 0$. Assume that $f \in D^s_{p,q}(A)$ with $s > \frac{1}{2} + \frac{1}{p'}$, $1 \leq p \leq 2$ and $1 \leq q \leq 2$. Then, as $n \to +\infty$*

$$\sup_{f \in D^s_{p,q}(A)} \mathbb{E}\|\hat{f}_n - f\|^2 \leq \mathcal{O}\left(\frac{n}{\log(n)} \right)^{-\frac{2s}{2s+1}}.$$

Minimax rates of convergence for density estimation over Besov spaces has been studied in detail by Donoho *et al* (1996). Hence, Theorem 4.1 shows that $\hat{f}_n$ is an adaptive estimator which converges to f with the minimax rate up to logarithmic factor for the problem of direct density estimation over $D^s_{p,q}(A)$. The extra logarithmic factor is usually called the price to pay for adaptivity to the unknown smoothness s.

4.2 The case of density deconvolution

Consider now the problem of density deconvolution under the Assumption 2.1 of ordinary smooth deconvolution.

Theorem 4.2 *Assume that the conditions of Theorem 3.2 hold $\eta = 1/2$ and $\alpha = 0$. Assume that $f \in D^s_{p,q}(A)$ with $s > \frac{1}{2} + \frac{1}{p'}$, $1 \leq p \leq 2$ and $1 \leq q \leq 2$. If $\nu(2-p) < ps^*$ then as $n \to +\infty$*

$$\sup_{f \in D^s_{p,q}(A)} \mathbb{E}\|\hat{f}_n - f\|^2 \leq \mathcal{O}\left(\frac{n}{\log(n)}\right)^{-\frac{2s}{2s+2\nu+1}},$$

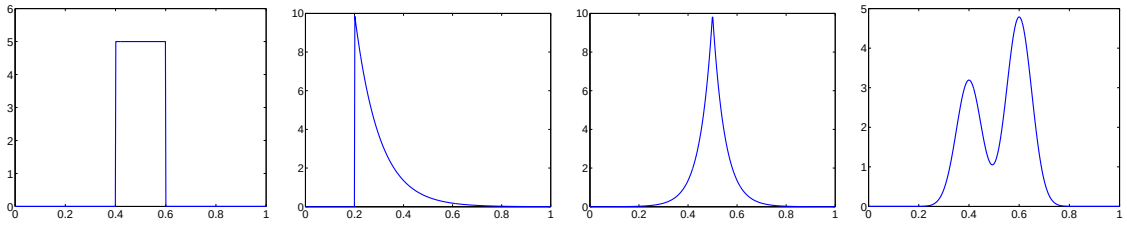
and if $\nu(2-p) \geq ps^*$ then as $n \to +\infty$

$$\sup_{f \in D^s_{p,q}(A)} \mathbb{E}\|\hat{f}_n - f\|^2 \leq \mathcal{O}\left(\frac{n}{\log(n)}\right)^{-\frac{2s^*}{2s^*+2\nu}}$$

Theorem 4.2 show that there is two different rates of convergence depending on whether $\nu(2-p) < ps^*$ or $\nu(2-p) \geq ps^*$. These two conditions are respectively referred to as the dense case when the worst functions f to estimate are spread uniformly over $[0,1]$, or the sparse case when the hardest functions to estimate have only one non-vanishing wavelet coefficient. This change in the rate of convergence, usually referred to as an elbow effect, has been studied in detail in nonparametric deconvolution problems in the white noise model by Johnstone *et al* (2004) and also Pensky and Sapatinas (2008). Theorem 4.2 shows that the rate of convergence of $\hat{f}_n$ corresponds to minimax rates (up to a logarithmic term) that have been obtained in related deconvolution problems either for density estimation or nonparametric regression in the white noise model (see e.g. Pensky and Vidakovic (1999), Fan and Koo (2002), Johnstone *et al* (2004), Pensky and Sapatinas (2008), and references therein).

5 Simulations

Simulations use the wavelet toolbox *Wavelab* of Matlab (Buckheit *et al*, 1995) and the fast algorithm for Meyer wavelet decomposition developed by Kolaczyk (1994). As in the simulation study of Bigot and Van Bellegem (2009), four test densities are considered: *Uniform distribution*: $f(x) = 5\mathbb{1}_{[0.4,0.6]}(x)$, *Exponential distribution*: $f(x) = 10e^{-10(x-0.2)}\mathbb{1}_{[0.2,+\infty[}(x)$, *Laplace distribution*: $f(x) = 10e^{-20|x-0.5|}$, and *MixtGauss distribution (mixture of two Gaussian variables)*: $X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$ with $\pi_1 = 0.4, \pi_1 = 0.6$, $\mu_1 = 0.4, \mu_2 = 0.6$ and $\sigma_1 = \sigma_2 = 0.05$. These four densities, displayed in Figure 1, exhibit different types of smoothness: the Uniform density is a piecewise constant function with two jumps, the Exponential distribution is a piecewise smooth function with a single jump, the Laplace density is a continuous function with a cusp at $x = 0.5$, whereas the MixtGauss density is infinitely differentiable.



(a) (b) (c) (d)

Figure 1: Test densities: (a) Uniform, (b) Exponential, (c) Laplace, (d) MixtGauss.

5.1 Direct density estimation

Assume that an i.i.d sample of variables $X_1, \dots, X_n$ is drawn from one the test densities shown in Figure 1. The empirical Fourier coefficients $\hat{f}_\ell = \frac{1}{n}\sum_{m=1}^{n} \exp(-2\pi i \ell X_m)$ are computed for $\ell = -N/2+1, \dots, N/2$, where $N = 2^J$ is a dyadic integer with J chosen such that $N \geq n$. The $\hat{f}_\ell$'s are then used as an input of the efficient algorithm of Kolaczyk (1994) in order to compute empirical scaling and wavelet coefficients $(\hat{c}_{j_0,k})_{k=0,\dots,2^{j_0-1}}, (\hat{\beta}_{j,k})_{k=0,\dots,2^j-1, j=j_0,\dots,J-1}$. This algorithm only requires $\mathcal{O}(N(\log(N))^2)$ operations to compute the empirical wavelet coefficients from a sample of Fourier coefficients of size N. According to Theorem 3.1, one can take $j_0 = \lfloor \log_2(\log(n)) \rfloor + 1$. Then, two hyperparameters essentially controls the quality of the estimator: the high-frequency cut-off parameter j_1 and the constant δ used in the definition of the thresholds $\tau_{j,k}$ given by equation (2.8). If the level j_1 is such that $2^{j_1} > n^\eta (\log n)^\alpha \geq 2^{j_1-1}$ for some $\eta > 0$, $\alpha \geq 0$, then one must take $\delta > \eta$. Increasing α deteriorates the rate of convergence of the additive term in the oracle inequality (3.1), so one can choose to set $\alpha = 0$. Following the choice made for the asymptotic study of $\hat{f}_n$ in Section 4, one can take $\eta = 1/2$ and according to Theorem 3.1 one should then take $\delta > 1/2$. However, it is not clear if a value for δ lower than $1/2$ would deteriorate or improve the quality of the estimator $\hat{f}_n$.

Alternatively, assume that j_1 is given. Then, another possibility for choosing δ is to take $\eta^* = (j_1 - 1)/\log_2(n)$ which is the smallest constant η that satisfies $2^{j_1} > n^\eta \geq 2^{j_1-1}$, and then take $\delta > \eta^*$. Combining the above arguments, we finally suggest to take

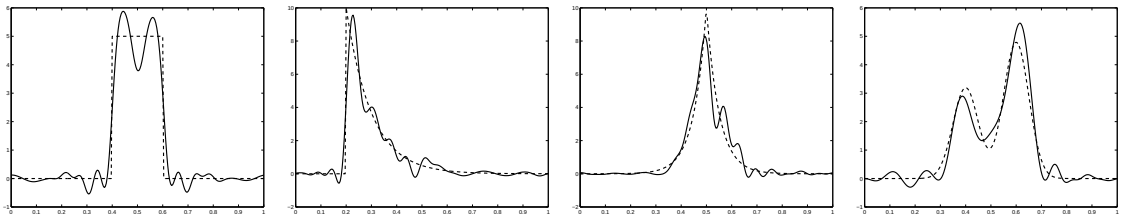
$$j_1 = j_1^* = \lfloor \frac{1}{2}\log_2(n) \rfloor + 1 \text{ and } \delta^* = (j_1 - 1)/\log_2(n). \tag{5.1}$$

To give an idea of the quality of $\hat{f}_n$, a typical example of estimation with $n = 200$, $j_1^* = 4$ and $\delta^* \approx 0.3925$ is given in Figure 2. Another possibility for choosing a smaller value for δ is to take $\alpha \neq 0$, and then to choose $\delta^* = \eta^* = (j_1 - 1 - \alpha \log_2(\log(n)))/\log_2(n)$ since η^* is the smallest constant η that satisfies $2^{j_1} > n^\eta \log(n)^\alpha \geq 2^{j_1-1}$. For $n = 200$, $\alpha = 0.5$ and $j_1 = 4$ this yields to the choice $\delta \approx 0.2351$. However, as already remarked, the oracle inequality (3.1) shows that taking $\alpha \neq 0$ may deteriorate the risk of $\hat{f}_n$.

The goal of this numerical section is thus to validate the above choices (5.1) for j_1 and δ, by studying the risk of $\hat{f}_n$ (compared to the risk of the oracle $\tilde{f}_n$) as a function of these two hyperparameters. More precisely, given j_1 and δ, we define

$$R_n(j_1, \delta) = \frac{\sum_{k=0}^{2^{j_0}-1}(\hat{c}_{j_0,k} - c_{j_0,k})^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}(\hat{\beta}_{j,k}\mathbb{1}_{\{|\hat{\beta}_{j,k}| \geqslant \tau_{j,k}\}} - \beta_{j,k})^2 + \sum_{j=j_1+1}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2}{\sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{j,k}^2, \sigma_{j,k}^2) + \sum_{j=j_1+1}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2} \tag{5.2}$$

To illustrate the usefulness of taking random thresholds $\tau_{j,k}$ depending of the location (j,k), we compare $R_n(j_1, \delta)$ with the risk $\tilde{R}_n(j_1, \delta)$ of the wavelet estimator obtained by taking a level-dependent threshold



(a) (b) (c) (d)

Figure 2: Typical reconstructions from a single simulation with $n = 200$, $j_1 = j_1^* = 4$ and $\delta = (j_1 - 1)/\log_2(n) \approx 0.3925$: (a) Uniform, (b) Exponential, (c) Laplace, (d) MixtGauss.

of the form $\delta\sqrt{j/n}$, as originally proposed by Donoho *et al* (1996), where

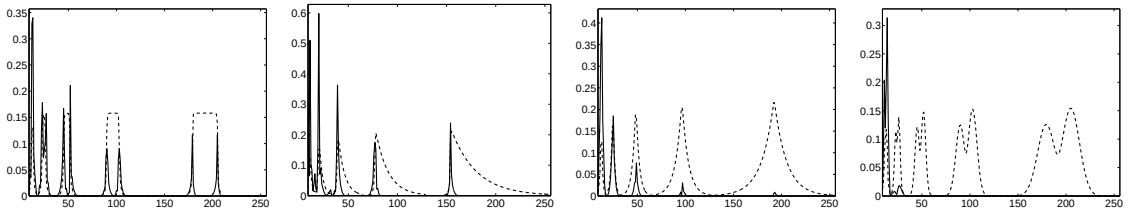
$$\tilde{R}_n(j_1,\delta) = \frac{\sum_{k=0}^{2^{j_0}-1}(\hat{c}_{j_0,k} - c_{j_0,k})^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}(\hat{\beta}_{j,k}\mathbb{1}_{\{|\hat{\beta}_{j,k}|\geqslant\delta\sqrt{j/n}\}} - \beta_{j,k})^2 + \sum_{j=j_1+1}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2}{\sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2 + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{j,k}^2,\sigma_{j,k}^2) + \sum_{j=j_1+1}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}^2}$$

Then, for each test density, $M = 100$ independent samples of size $n = 100, 200$ are drawn. Empirical average of $R_n(j_1,\delta)$ and $\tilde{R}_n(j_1,\delta)$ over these M repetitions are plotted in Figure 4 ($n = 100$) and Figure 5 ($n = 200$) with $\delta \in [0,5]$ and $j_1^* \leq j_1 \leq j_1^* + 2$. For $n = 200$ and $N = 256$, we also display in Figure 3 the true wavelet coefficients $\beta_{j,k}$ and the standard deviation $\sigma_{j,k}$ to show the ideal thresholding performed by the oracle estimator $\tilde{f}_n$. For the Uniform, Exponential and Laplace distributions and $j_1 = 4$, it can be seen that $|\beta_{j,k}| \geq \sigma_{j,k}$ for all $j_0 \leq j \leq j_1$ and $0 \leq k \leq 2^j - 1$, which means that the oracle estimator $\tilde{f}_n$ performs no thresholding and keeps all empirical wavelet coefficients $\beta_{j,k}$ for $j \leq j_1$. For the MixtGauss distribution and $j_1 = 4$ the behavior of the oracle estimator is different as for some $0 \leq k \leq 2^{j_1} - 1$, one can observe that $|\beta_{j,k}| < \sigma_{j,k}$. One retrieves this behavior in the first column of Figure 5 (case $j_1 = 4$) for the curves $R_n(j_1,\delta)$ and $\tilde{R}_n(j_1,\delta)$ which both have a minimum at $\delta = 0$ (no thresholding is done) for the Uniform, Exponential and Laplace distributions. For the MixtGauss distribution, $R_n(j_1,\delta)$ has a minimum at $\delta \approx 0.4$ while $\tilde{R}_n(j_1,\delta)$ has a minimum at $\delta \approx 0.8$. For $5 \leq j_1 \leq 7$ and the Uniform, Laplace and MixtGauss distributions, $R_n(j_1,\delta)$ has a minimum at some $\delta \in [0.4,1]$ and the value of $R_n(j_1,\delta)$ at this point is smaller that the minimum of $\tilde{R}_n(j_1,\delta)$. This indicates that taking thresholds depending on the location (j,k) can improve the quality of the estimation. For the Exponential distribution, the estimator with a level-dependent threshold of the form $\delta\sqrt{j/n}$ performs generally better than $\hat{f}_n$. Similar comments can be made for the behavior of the curves displayed in Figure 4 (case $n = 100$).

Moreover, it can been that the smallest value of the risk of $\hat{f}_n$ relative to the risk of the oracle $\tilde{f}_n$ are obtained for $j_1 = j_1^* = \lfloor \frac{1}{2}\log_2(n) \rfloor + 1$. This indicates that introducing wavelet coefficients at resolution level larger than j_1^* generally deteriorates the quality of the estimation. Finally, note that the curves in Figure 4 and Figure 5 tends to confirm that the choice (5.1) for j_1 and δ is reasonable, and leads to very satisfactory estimators.

5.2 Density deconvolution

In the case of density deconvolution, we propose to compare the performances of our wavelet approach with those of the adaptive density deconvolution estimator of Comte, Rozenholc and Taupin (2006a), Comte *et al* (2006b) that is based on penalized contrast minimization over a collection of models containing square integrable functions with Fourier transform having a compact support. Such an estimator is therefore a band-limited function, see Comte *et al* (2006b) for further details. Moreover, this method can be viewed as a kind of adaptive linear wavelet estimator with a Shannon wavelet



(a) (b) (c) (d)

Figure 3: True wavelet coefficients $|\beta_{j,k}|$ (solid curve), and standard deviation $\sigma_{j,k}$ (dashed curves) as a function of $2^j + k$ for $j = j_0, \ldots, J-1, k = 0, \ldots, 2^j - 1$ (with $j_0 = 3$, $J = 8$), for each test density: (a) Uniform, (b) Exponential, (c) Laplace, (d) MixtGauss.

basis which is also a band-limited function like the Meyer wavelet but less localized in the time domain. Comte *et al* (2006b) have shown that the model selection procedure performs very well for finite samples, compared with other standard estimators. In particular, this estimator outperforms the kernel estimator, even when the bandwidth parameter is selected in a data-driven way which makes this procedure as the most challenging competitor in our simulations.

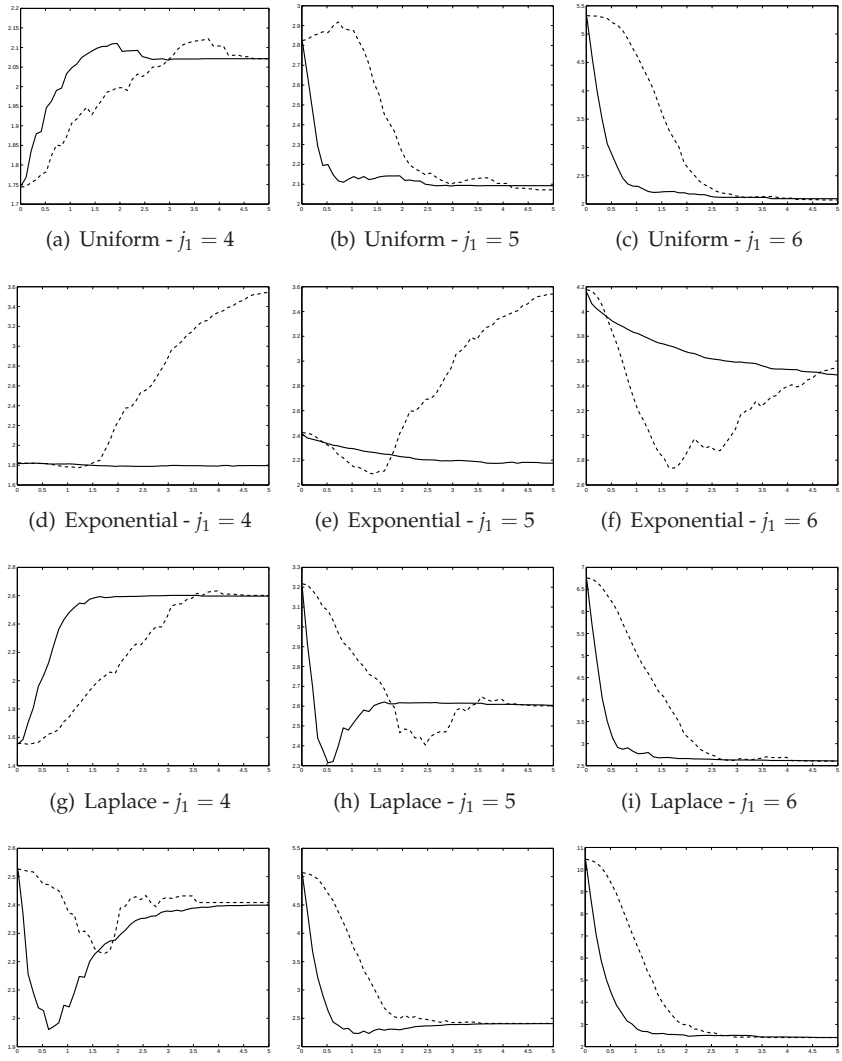
Observations $Y_i, i = 1, \ldots, n$ are generated from the additive model $Y_i = X_i + \epsilon_i$, where X_i are independent realizations from one of the test functions f displayed in Figure 1, and the ϵ_i's are i.d.d. additive errors with density h. Results are presented with a Laplace measurement error, that is $h(x) = (\sqrt{2}\sigma_\epsilon)^{-1} \exp(-\sqrt{2}|x|/\sigma_\epsilon)$, for $x \in \mathbb{R}$ and where σ_ϵ is the standard deviation of the errors. The Fourier coefficients of h are given by $h_\ell = (1 + 2\sigma_\epsilon^2 \pi^2 \ell^2)^{-1}$, $\ell \in \mathbb{Z}$ which thus corresponds to the case of ordinary smooth deconvolution with $\nu = 2$. The main quantities in the simulations are the sample size n and the root signal-to-noise ratio defined by $s2n = \sigma_X / \sigma_\epsilon$ with $\sigma_X = \sqrt{\mathrm{Var}(X_1)}$.

According to Theorem 3.2, one can take $j_0 = \lfloor \log_2(\log(n)) \rfloor + 1$ and important quantities to control the quality of estimation by wavelet thresholding are the highest resolution level j_1 and the tuning parameter δ. If j_1 is such that $2^{j_1} > n^{\eta/(\nu+1)} (\log n)^\alpha \geq 2^{j_1 - 1}$ for some $\eta > 0$, $\alpha \geq 0$, then Theorem 3.2 suggests to take $\delta = \eta \left(1 + \frac{\nu}{\nu+1}\right)$. As already remarked, for ill-posed inverse problems, a smaller j_1 than in the direct case should be used. However, for $n = 200$, following the asymptotic considerations in Section 4, the choices $\eta = 1/2$, $\alpha = 0$ and $\nu = 2$ yield to $j_1 = \lfloor \frac{1}{6} \log_2(n) \rfloor + 1 = 2$ which is smaller than $j_0 = \lfloor \log_2(\log(n)) \rfloor + 1 = 3$. Hence, setting in advance values for η and α may yield a theoretical choice for j_1 that cannot be used in practice. Note that this issue has been noticed in several papers on deconvolution by wavelet thresholding, see Johnstone *et al* (2004), Bigot and Van Bellegem (2009).

Alternatively, let us argue as in the direct case, by assuming that j_1 is given. If one sets $\alpha = 0$, then for choosing δ one can take $\eta^* = (\nu+1)(j_1 - 1)/\log_2(n)$ which is the smallest constant η that satisfies $2^{j_1} > n^{\eta/(\nu+1)} \geq 2^{j_1 - 1}$, and then take $\delta = \left(1 + \frac{\nu}{\nu+1}\right) \eta^* = (2\nu+1)(j_1 - 1)/\log_2(n)$. A smaller value for δ can also be made, by choosing $\alpha \neq 0$ and by taking $\delta = \left(1 + \frac{\nu}{\nu+1}\right) \eta^*$ with $\eta^* = (\nu+1)(j_1 - 1 - \alpha \log_2(\log(n)))/\log_2(n)$.

We report results for $n = 100, 200$ and $s2n = 3$, which a relatively large signal-to-noise ratio. To give an idea of the quality of $\hat{f}_n$ and to compare it with the model selection estimator, a typical example of estimation with $n = 200$, $j_1 = j_0 = 3$, $\alpha = 0.5$ and $\delta = (2\nu+1)(j_1 - 1 - \alpha \log_2(\log(n)))/\log_2(n) \approx 0.5215$ is given in Figure 6. Both methods perform similarly for the estimation of the smooth density MixtGauss. For the three non-smooth densities, wavelet thresholding performs much better than the model selection estimator. This comparison on a single simulation tends to confirm the superiority of wavelet-based methods over those based on Fourier decompositions for the reconstruction of signals with local singularities.

However, it is not clear how to choose j_1 in practice. Indeed, an optimal theoretical level can be



(a) Uniform - $j_1 = 4$ (b) Uniform - $j_1 = 5$ (c) Uniform - $j_1 = 6$

(d) Exponential - $j_1 = 4$ (e) Exponential - $j_1 = 5$ (f) Exponential - $j_1 = 6$

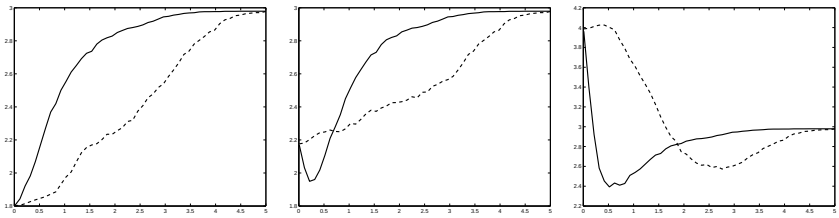
(g) Laplace - $j_1 = 4$ (h) Laplace - $j_1 = 5$ (i) Laplace - $j_1 = 6$

(j) MixtGauss - $j_1 = 4$ (k) MixtGauss - $j_1 = 5$ (l) MixtGauss - $j_1 = 6$

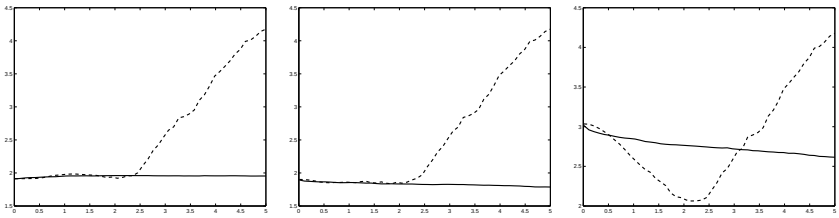
Figure 4: Direct density estimation with $n = 100$. Evolution of $R_n(j_1, \delta)$ (solid curves) and $\tilde{R}_n(j_1, \delta)$ (dashed curves) as a function of $\delta \in [0, 5]$ for different values of $j_1 \geq j_1^* = \lfloor \frac{1}{2} \log_2(n) \rfloor + 1 = 4$

too small, but taking a too high level of resolution may introduce some instability in the estimation. Moreover, it is of interested to study if δ can be smaller that $\eta \left(1 + \frac{\nu}{\nu+1}\right)$. A goal of the following simulation study is thus to identify a reasonable empirical range of values for j_1 and δ.

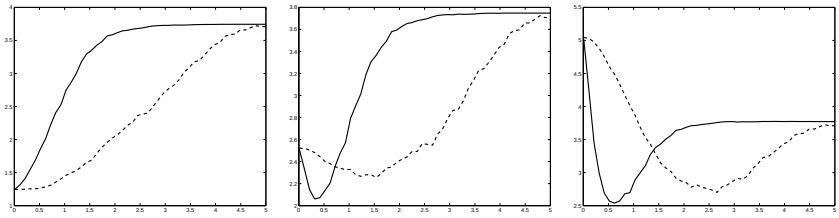
For a given j_1 and δ, the risk $R_n(j_1, \delta)$ of the wavelet-based estimator, see (5.2), will be compared to the risk of the model section procedure divided by the risk of the oracle $\tilde{f}_n$. For each test density, $M = 100$ independent samples of size $n = 100, 200$ are drawn for $s2n = 3$. Empirical average of $R_n(j_1, \delta)$ and of the risk of the model selection estimator (divided by the oracle risk) over these M repetitions are displayed in Figure 7 ($n = 100$) and Figure 8 ($n = 200$) for $\delta \in [0, 5]$ and $j_0 \leq j_1 \leq j_0 + 2$. For $j_1 = j_0$, wavelet thresholding clearly outperforms the model selection estimator for all values of δ and all densities, expect for the Uniform distribution with $n = 200$ for which it can be seen that model selection is slightly better if δ is larger than 0.2. These simulations, also show that the choice $j_1 = j_0 = 3$



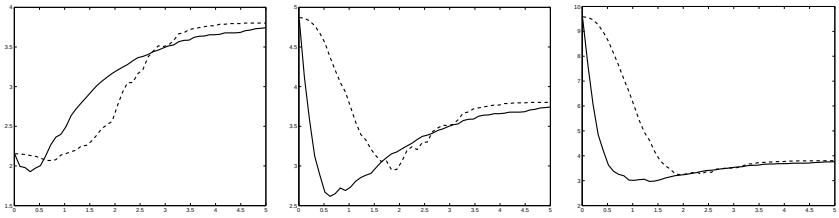
(a) Uniform - $j_1 = 4$ (b) Uniform - $j_1 = 5$ (c) Uniform - $j_1 = 6$



(d) Exponential - $j_1 = 4$ (e) Exponential - $j_1 = 5$ (f) Exponential - $j_1 = 6$



(g) Laplace - $j_1 = 4$ (h) Laplace - $j_1 = 5$ (i) Laplace - $j_1 = 6$



(j) MixtGauss - $j_1 = 4$ (k) MixtGauss - $j_1 = 5$ (l) MixtGauss - $j_1 = 6$

Figure 5: Direct density estimation with $n = 200$. Evolution of $R_n(j_1, \delta)$ (solid curves) and $\tilde{R}_n(j_1, \delta)$ (dashed curves) as a function of $\delta \in [0, 5]$ for different values of $j_1 \geq j_1^* = \lfloor \frac{1}{2} \log_2(n) \rfloor + 1 = 4$.

yields the best results. This observation is consistent with the condition of Theorem 3.2 that suggests a smaller j_1 for ill-posed inverse problems than in the direct case. It also confirms that introducing a higher level of resolution clearly deteriorates the quality of the estimator when compared to the oracle risk.

Combining the above remarks, we finally suggest the following choice

$$j_1 = j_0 = \lfloor \log_2(\log(n)) \rfloor + 1 \text{ and } \delta = (2\nu + 1)(j_1 - 1 - \alpha \log_2(\log(n)))/\log_2(n) \text{ with } \alpha = 0.5.$$

For $n = 100$, this yields to $j_1 = 3, \delta \approx 0.6761$ and for $n = 200$ to $j_1 = 3, \delta \approx 0.5215$. The curves in Figure 7 and Figure 8 indicate that such a choice is reasonable, and leads to satisfactory estimators.

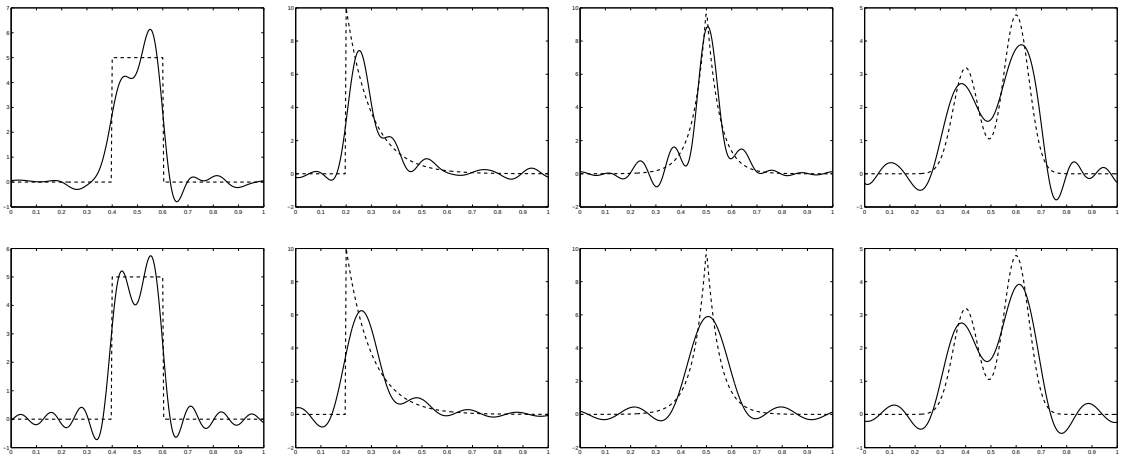


Figure 6: Typical reconstructions from a single simulation with $n = 200$ for the four test densities Uniform (1st column), Exponential (2nd column), Laplace (3rd column), MixtGauss (4th column) contaminated with Laplace noise, by wavelet thresholding (1st row) with $j_1 = j_0 = 3$, $\delta \approx 0.5215$, and model selection (2nd row).

Appendix

In what follows C will denote a generic constant whose value may change from line to line. Proofs are given for the case where h satisfies Assumption 2.1 and Assumption 3.1. The proofs for the case of direct density estimation follow from the same arguments.

5.3 Technical lemmas

We start this technical section by a set of lemmas that will be used in the proof of Theorem 3.2. For all the results presented below, it is supposed that $f \in D^2([0,1])$ with $\|f\|_\infty < +\infty$.

To prove these results we will use the following properties which come directly from the fact that Meyer wavelets are band-limited and that under Assumption 2.1 $|h_\ell| \sim |\ell|^{-\nu}$.

$$|\psi_{j,k}| \leq 2^{-j/2} \text{ and } \#C_j \leq 4\pi 2^j \tag{5.3}$$

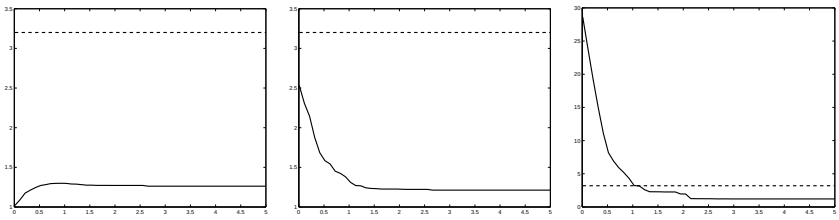
$$|h_\ell|^{-2} \leq C 2^{2j\nu} \text{ for all } \ell \in C_j, \tag{5.4}$$

where $\#C_j$ denotes the cardinality of the set C_j.

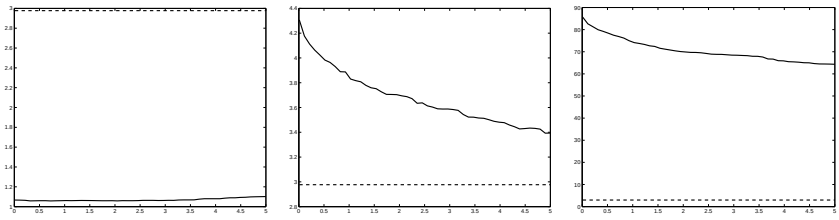
Lemma 5.1 *For all integer j, $\sigma_{j,k}^2 = \mathbb{E}(\hat{\beta}_{j,k} - \beta_{j,k})^2 \leq V_{j,k} \leq C\|f\|_\infty \frac{2^{2j\nu}}{n}$ and $\sigma_{j_0,k}^2 = \mathbb{E}(\hat{c}_{j_0,k} - c_{j_0,k})^2 \leq C\|f\|_\infty \frac{2^{2j_0\nu}}{n}$.*

Proof: recall that $V_{j,k}$ is an upper bound for $\sigma_{j,k}^2$, and that from equation (2.5), one has that $V_{j,k} = \frac{1}{n}\int_{\mathbb{R}} |\tilde{\psi}_{j,k}(y)|^2 f^Y(y)dy$. Then, remark that since h satisfies Assumption 3.1, it holds that

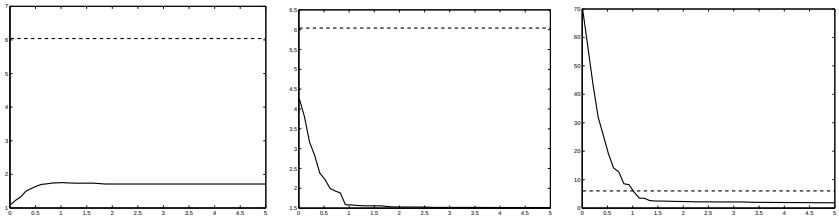
$$\begin{aligned} V_{j,k} &= \frac{1}{n}\sum_{m\in\mathbb{Z}} \int_m^{m+1} |\tilde{\psi}_{j,k}(y)|^2 f^Y(y)dy = \frac{1}{n}\sum_{m\in\mathbb{Z}} \int_m^{m+1} |\tilde{\psi}_{j,k}(y)|^2 \int_0^1 f(u)h(u-y)dudy \\ &\leq C\|f\|_\infty \frac{1}{n}\sum_{m\in\mathbb{Z}} \int_m^{m+1} |\tilde{\psi}_{j,k}(y)|^2 \int_0^1 \frac{1}{|u-y|^\gamma} dudy \end{aligned}$$



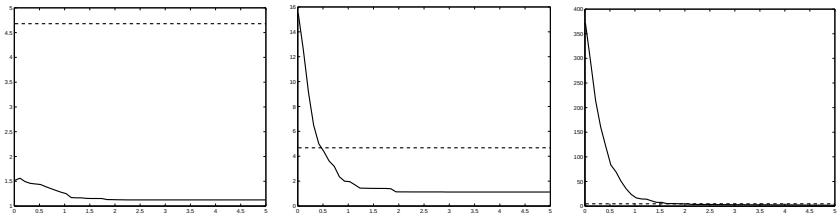
(a) Uniform - $j_1 = 3$ (b) Uniform - $j_1 = 4$ (c) Uniform - $j_1 = 5$



(d) Exponential - $j_1 = 3$ (e) Exponential - $j_1 = 4$ (f) Exponential - $j_1 = 5$



(g) Laplace - $j_1 = 3$ (h) Laplace - $j_1 = 4$ (i) Laplace - $j_1 = 5$



(j) MixtGauss - $j_1 = 3$ (k) MixtGauss - $j_1 = 4$ (l) MixtGauss - $j_1 = 5$

Figure 7: Density deconvolution with $s2n = 3$ and $n = 100$. Evolution of $R_n(j_1, \delta)$ as a function of $\delta \in [0, 5]$ for different values of $j_1 \geq j_0 = 3$ (solid curves). The dotted lines represent the risk of the model selection estimator divided by the risk of the oracle.

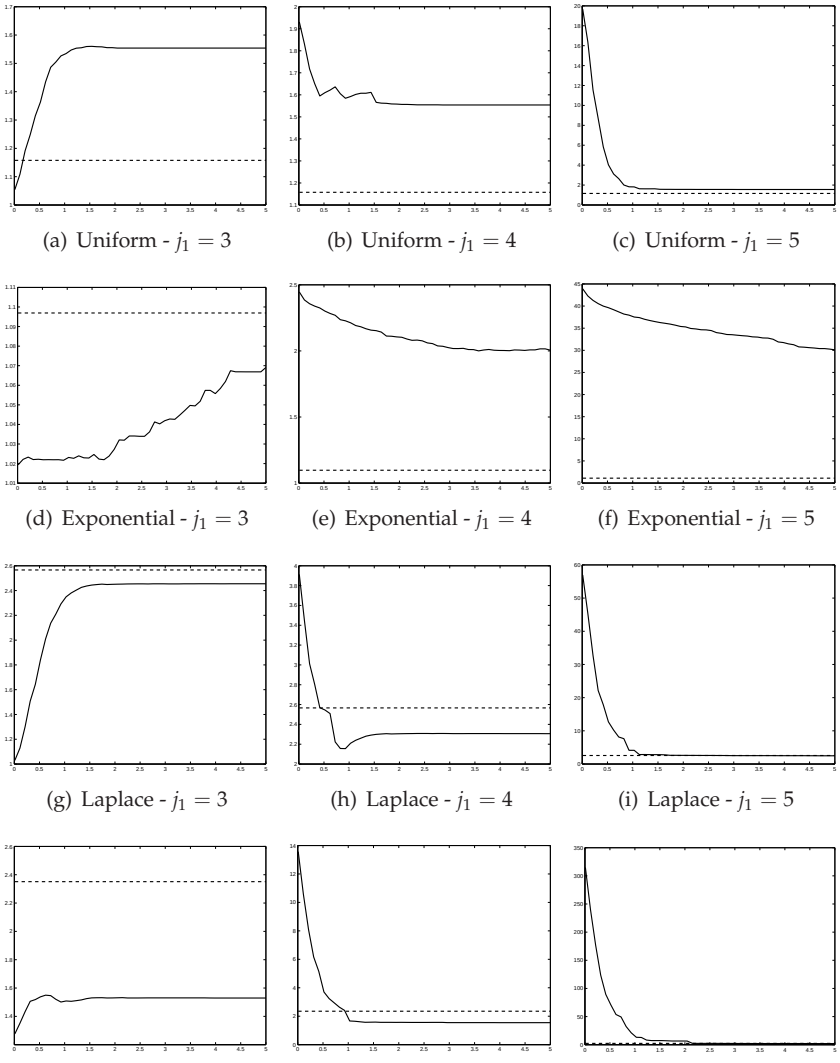
where $\gamma > 1$ is the real defined in Assumption 3.1. Now, since $\tilde{\psi}_{j,k}(y)$ is a periodic function on $\mathbb{R}$ with period 1 it follows that

$$V_{j,k} \leq C\|f\|_\infty \frac{1}{n} \int_0^1 |\tilde{\psi}_{j,k}(y)|^2 dy \sum_{m \in \mathbb{Z}} |m|^{-\gamma}$$

Finally, using Parseval relation and the bounds (5.3) and (5.4), it follows that $\int_0^1 |\tilde{\psi}_{j,k}(y)|^2 dy = \sum_{\ell \in C_j} |\tilde{\psi}_\ell^{j,k}|^2 = \sum_{\ell \in C_j} \frac{|\psi_\ell^{j,k}|^2}{|h_\ell|^2} \leq C 2^{2j\nu}$ which finally implies (using the fact that $\gamma > 1$,)

$$V_{j,k} \leq C\|f\|_\infty \frac{2^{2j\nu}}{n}$$

which completes the proof. The argument is the same to bound $\sigma^2_{j_0,k}$. □



(a) Uniform - $j_1 = 3$ (b) Uniform - $j_1 = 4$ (c) Uniform - $j_1 = 5$

(d) Exponential - $j_1 = 3$ (e) Exponential - $j_1 = 4$ (f) Exponential - $j_1 = 5$

(g) Laplace - $j_1 = 3$ (h) Laplace - $j_1 = 4$ (i) Laplace - $j_1 = 5$

(j) MixtGauss - $j_1 = 3$ (k) MixtGauss - $j_1 = 4$ (l) MixtGauss - $j_1 = 5$

Figure 8: Density deconvolution with $s2n = 3$ and $n = 200$. Evolution of $R_n(j_1, \delta)$ as a function of $\delta \in [0, 5]$ for different values of $j_1 \geq j_0 = 3$ (solid curves). The dotted lines represent the risk of the model selection estimator divided by the risk of the oracle.

Lemma 5.2 *For all* $p \geq 2$, $\mathbb{E}(\hat{\beta}_{j,k} - \beta_{j,k})^{2p} \leq C \max(\|f\|_\infty^p, 1) \left(\frac{2^{2jvp}}{n^p} + \frac{2^{jp(2v+1)}}{n^{2p-1}} \right)$.

Proof: by definition $\hat{\beta}_{j,k} - \beta_{j,k} = \frac{1}{n} \sum_{m=1}^n Z_m$ where $Z_m = \sum_{\ell \in C_j} \tilde{\psi}_\ell^{j,k} \left(e^{-2\pi i \ell Y_m} - f_\ell h_\ell \right)$. Remark that for all m, $\mathbb{E} Z_m = 0$, and that from Lemma 5.1 $\mathrm{Var}(Z_m) \leq C \|f\|_\infty 2^{2jv}$. Since f and h are densities, $|f_\ell h_\ell| \leq 1$. Hence using (5.3) and (5.4), this implies that $|Z_m| \leq 2 \sum_{\ell \in C_j} \frac{|\psi_\ell^{j,k}|}{|h_\ell|} \leq C 2^{j(v+1/2)}$. Then the result follows from Rosenthal's inequality (see Rosenthal (1972)). □

Lemma 5.3 *For any positive* x, $\mathbb{P}\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \sqrt{2 V_{jk} x} + \frac{\eta_j}{3n} x \right) \leq 2 \exp(-x)$, *where* $\eta_j = \sum_{\ell \in C_j} |\tilde{\psi}_{jk}^\ell|$.

Proof: note that $\hat{\beta}_{j,k} - \beta_{j,k} = \frac{1}{n} \sum_{m=1}^n (W_m - \mathbb{E} W_m)$ where $W_m = \sum_{\ell \in C_j} \tilde{\psi}_\ell^{j,k} e^{-2\pi i \ell Y_m}$. Then, remark that

by definition $V_{jk} = \frac{1}{n^2}\sum_{m=1}^{n} \mathbb{E}|W_m|^2$, and that $|W_m| \leq \sum_{\ell \in C_j} |\tilde{\psi}_{jk}^{\ell}|$. Hence, the Z_m's are bounded random variables, and thus the result follows from Bernstein's inequality (see e.g. Proposition 2.9 in Massart (2006)). □

Lemma 5.4 *For any positive x,*

$$\mathbb{P}\left(V_{jk} \geq \hat{V}_{jk} + \sqrt{2\frac{\eta_j^2}{n^2}\hat{V}_{jk}x} + \kappa\frac{\eta_j^2}{n^2}x \right) \leq \exp(-x).$$

where $\kappa = \frac{4}{3} + \sqrt{\frac{5}{3}}$ and $\eta_j = \sum_{\ell \in C_j} |\tilde{\psi}_{jk}^{\ell}|$.

Proof: the proof is inspired by the proof of Lemma 1 in Reynaud-Bouret and Rivoirard (2008). By definition $\hat{V}_{jk} = \frac{1}{n^2}\sum_{m=1}^{n} W_m$ with $W_m = \sum_{\ell,\ell' \in C_j} \tilde{\psi}_{\ell}^{j,k}\overline{\tilde{\psi}_{\ell'}^{j,k}} e^{-2\pi i(\ell-\ell')Y_m}$. Then, remark that $|W_m| \leq \sum_{\ell,\ell' \in C_j} |\tilde{\psi}_{\ell}^{j,k}||\tilde{\psi}_{\ell'}^{j,k}|$ which implies that $|W_m| \leq \eta_j^2$ for all $m = 1, \ldots, n$. Moreover, one can remark that $\mathbb{E}|W_m|^2 = \mathbb{E}|\tilde{\psi}_{j,k}(Y_m)|^4 = \int_{\mathbb{R}} |\tilde{\psi}_{j,k}(y)|^4 f^Y(y)dy$. Then, since $|\tilde{\psi}_{j,k}(y)|^2 \leq \left(\sum_{\ell \in C_j} |\tilde{\psi}_{\ell}^{j,k}|\right)^2$, it follows that

$$\mathbb{E}|W_m|^2 \leq \eta_j^2 \int_{\mathbb{R}} |\tilde{\psi}_{j,k}(y)|^2 f^Y(y)dy \leq \eta_j^2 n V_{jk}.$$

Hence, by applying Bernstein's inequality (see e.g. Proposition 2.9 in Massart (2006)), one obtains that for any positive x

$$\mathbb{P}\left(V_{jk} \geq \hat{V}_{jk} + \sqrt{\frac{2\eta_j^2}{n^2}V_{jk}x} + \frac{\eta_j^2}{3n^2}x \right) \leq \exp(-x). \tag{5.5}$$

Now, let $u = \frac{\eta_j^2}{n^2}x$ and let $g(y) = y^2 - \sqrt{2u}y - \frac{u}{3}$ for $y \geq 0$. From (5.5), one has that

$$\mathbb{P}\left(g(\sqrt{V_{jk}}) \geq \hat{V}_{jk} \right) \leq \exp(-x).$$

As V_{jk} and $\hat{V}_{jk}$ are positive, one can check that it is possible to invert the inequality $g(\sqrt{V_{jk}}) \geq V_{jk}$ to obtain that

$$\mathbb{P}\left(\sqrt{V_{jk}} \geq g^{-1}(\hat{V}_{jk}) \right) \leq \exp(-x),$$

and that $g^{-1}(y) = \sqrt{y + \frac{5u}{6}} + \sqrt{\frac{u}{2}}$. Hence, one obtains that for any positive x

$$\mathbb{P}\left(V_{jk} \geq \hat{V}_{jk} + \frac{4}{3}u + \sqrt{2u\hat{V}_{jk} + \frac{5u^2}{3}} \right) \leq \exp(-x).$$

Now, using the fact that $\sqrt{2u\hat{V}_{jk} + \frac{5u^2}{3}} \leq \sqrt{2u\hat{V}_{jk}} + \sqrt{\frac{5u^2}{3}}$, it follows that

$$\hat{V}_{jk} + \frac{4}{3}u + \sqrt{2u\hat{V}_{jk} + \frac{5u^2}{3}} \leq \hat{V}_{jk} + \sqrt{2u\hat{V}_{jk}} + \kappa u,$$

which completes the proof. □

5.4 Proof of Theorem 3.2

Let us define the following set of integers $\Lambda_n = \{(j,k); j_0 \leq j \leq j_1, 0 \leq k \leq 2^j - 1\}$. We first establish the following proposition:

Proposition 5.1 *Let $\epsilon > 0$. Then, for any subset of indices $m \subset \Lambda_n$*

$$\begin{aligned}\|\hat{f}_n - f\|^2 \quad \leq \quad & \sum_{k=0}^{2^{j_0}-1} (\hat{c}_{j_0,k} - c_{j_0,k})^2 + \left(\frac{2+\epsilon}{\epsilon}\right)^2 \sum_{(j,k)\in\Lambda_n\setminus m} \beta_{jk}^2 + (2+\epsilon) \sum_{(j,k)\in m} \left(\hat{\beta}_{jk} - \beta_{jk}\right)^2 \\ & + \left(\frac{2+\epsilon}{\epsilon}\right) \sum_{(j,k)\in m} \tau_{jk}^2 + \left(\frac{2+\epsilon}{\epsilon}\right)^2 \sum_{j=j_1+1}^{+\infty} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \\ & + \left(\frac{2+\epsilon}{\epsilon}\right) \sum_{(j,k)\in\hat{m}} \left((1+\epsilon)\left(\hat{\beta}_{jk} - \beta_{jk}\right)^2 - \tau_{jk}^2\right)\end{aligned}$$

where $\hat{m} = \{(j,k); |\beta_{j,k}| \geqslant \tau_{j,k}\, , j_0 \leq j \leq j_1, 0 \leq k \leq 2^j - 1\}$.

Proof: the proof is inspired by model selection techniques (see Massart (2006) and also Reynaud-Bouret and Rivoirard (2008)). Let $m \subset \Lambda_n$ and define

$$\hat{f}_m = \sum_{(j,k)\in m} \hat{\beta}_{jk}\psi_{jk} \text{ and } f_m = \sum_{(j,k)\in m} \beta_{jk}\psi_{jk}.$$

Note that $\hat{f}_{\hat{m}}$ is given by $\hat{f}_{\hat{m}} = \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{jk} \mathbb{1}_{\{|\beta_{j,k}| \geqslant \tau_{j,k}\}} \psi_{j,k}$. Then, define

$$\gamma_m = -\sum_{(j,k)\in m} \hat{\beta}_{jk}^2 \text{ and } \text{pen}(m) = \sum_{(j,k)\in m} \tau_{jk}^2,$$

and one can check that

$$\hat{m} = \arg\min_{m\in\Lambda_n} \gamma_m + \text{pen}(m).$$

Now let $\tilde{f} = \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}$ and remark that for any $m \subset \Lambda_n$

$$\gamma_m = \|\hat{f}_m - \tilde{f}\|^2 - \sum_{j=j_0}^{+\infty} \sum_{k=0}^{2^j-1} \beta_{jk}^2 - 2 \sum_{(j,k)\in m} \hat{\beta}_{jk}(\hat{\beta}_{jk} - \beta_{jk}). \tag{5.6}$$

Then, by definition of $\hat{m}$, equality (5.6) implies that for all $m \subset \Lambda_n$

$$\|\hat{f}_{\hat{m}} - \tilde{f}\|^2 \leq \|\hat{f}_m - \tilde{f}\|^2 - 2 \sum_{(j,k)\in m} \hat{\beta}_{jk}(\hat{\beta}_{jk} - \beta_{jk}) + 2 \sum_{(j,k)\in\hat{m}} \hat{\beta}_{jk}(\hat{\beta}_{jk} - \beta_{jk}) + \text{pen}(m) - \text{pen}(\hat{m}).$$

Using the Pythagorean equality $\|\hat{f}_m - \tilde{f}\|^2 = \|f_m - \tilde{f}\|^2 + \|\hat{f}_m - f_m\|^2$ one finally obtains that

$$\begin{aligned}\|\hat{f}_{\hat{m}} - \tilde{f}\|^2 \quad \leq \quad & \|f_m - \tilde{f}\|^2 - \|\hat{f}_m - f_m\|^2 + 2 \sum_{(j,k)\in\hat{m}} \hat{\beta}_{jk}(\hat{\beta}_{jk} - \beta_{jk}) - 2 \sum_{(j,k)\in m} \beta_{jk}(\hat{\beta}_{jk} - \beta_{jk}) \\ & + \text{pen}(m) - \text{pen}(\hat{m}).\end{aligned} \tag{5.7}$$

Let $\epsilon > 0$ and remark that,

$$2 \sum_{(j,k)\in\hat{m}} \hat{\beta}_{jk}(\hat{\beta}_{jk} - \beta_{jk}) - 2 \sum_{(j,k)\in m} \beta_{jk}(\hat{\beta}_{jk} - \beta_{jk}) \leq 2\left(\|\hat{f}_{\hat{m}} - \tilde{f}\| + \|f_m - \tilde{f}\|\right) \sqrt{\sum_{(j,k)\in\hat{m}\cup m} (\hat{\beta}_{jk} - \beta_{jk})^2} \tag{5.8}$$

Then, remark that $\sum_{(j,k)\in\hat{m}\cup m}(\hat{\beta}_{jk}-\beta_{jk})^2 \leq \|\hat{f}_m - f_m\|^2 + \|\hat{f}_{\hat{m}} - f_{\hat{m}}\|^2$ and thus by using twice the inequality $2ab \leq \theta a^2 + (1/\theta)b^2$ with $\theta = 2/(2+\epsilon)$ and $\theta = 2/\epsilon$, and by inserting inequality (5.8) in (5.7), one obtains that

$$\begin{aligned}\frac{\epsilon}{2+\epsilon}\|\hat{f}_{\hat{m}} - \tilde{f}\|^2 &\leq \frac{2+\epsilon}{\epsilon}\|f_m - \tilde{f}\|^2 + \epsilon\|\hat{f}_m - f_m\|^2 + \mathrm{pen}(m) \\ &\quad +(1+\epsilon)\|\hat{f}_{\hat{m}} - f_{\hat{m}}\|^2 - \mathrm{pen}(\hat{m}). \end{aligned} \tag{5.9}$$

which completes the proof since $\hat{f}_n = \sum_{k=0}^{2^{j_0}-1}\hat{c}_{j_0k}\phi_{j_0,k} + \hat{f}_{\hat{m}}$. □

Next we prove the following lemma:

Lemma 5.5 *For any $\delta > \eta\left(1+\frac{\nu}{\nu+1}\right)$, there exists $\epsilon = \epsilon(\delta) > 0$ such that $\delta > (1+\epsilon)\left(1+\frac{\nu}{\nu+1}\right)$, and a positive constant C not depending on n such that*

$$\mathbb{E}\left(\sum_{(j,k)\in\hat{m}}(1+\epsilon)\left(\hat{\beta}_{jk}-\beta_{jk}\right)^2 - \tau_{jk}^2\right) \leq C\max(\|f\|_\infty, 1)\max(\log(n)^\alpha, 1)\frac{(\log n)^{\alpha(2\nu+1)}}{n}.$$

Proof: let $Z = \sum_{(j,k)\in\hat{m}}(1+\epsilon)\left(\hat{\beta}_{jk}-\beta_{jk}\right)^2 - \tau_{jk}^2$. By definition of $\hat{m}$

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\left((1+\epsilon)\left(\hat{\beta}_{jk}-\beta_{jk}\right)^2 - \tau_{jk}^2\right)\mathbb{1}_{\{|\hat{\beta}_{jk}|\geq\tau_{jk}\}} \\ &\leq \mathbb{E}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\left((1+\epsilon)\left(\hat{\beta}_{jk}-\beta_{jk}\right)^2\right)\mathbb{1}_{\{|\hat{\beta}_{jk}|\geq\tau_{jk}\cap|\hat{\beta}_{jk}-\beta_{jk}|\geq\frac{\tau_{jk}}{\sqrt{1+\epsilon}}\}} \\ &\leq (1+\epsilon)\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\mathbb{E}\left(\left(\hat{\beta}_{jk}-\beta_{jk}\right)^2\mathbb{1}_{\{|\hat{\beta}_{jk}-\beta_{jk}|\geq\frac{\tau_{jk}}{\sqrt{1+\epsilon}}\}}\right)\end{aligned}$$

Now by applying Holder inequality, one has that for all $p \geq 2$ and $q > 1$ such that $\frac{1}{p}+\frac{1}{q} = 1$

$$\mathbb{E}Z \leq (1+\epsilon)\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\left(\mathbb{E}\left(\hat{\beta}_{jk}-\beta_{jk}\right)^{2p}\right)^{1/p}\left(\mathbb{P}\left(|\hat{\beta}_{jk}-\beta_{jk}| \geq \frac{\tau_{jk}}{\sqrt{1+\epsilon}}\right)\right)^{1/q} \tag{5.10}$$

Now by Lemma 5.2 one has that

$$\left(\mathbb{E}\left(\hat{\beta}_{jk}-\beta_{jk}\right)^{2p}\right)^{1/p} \leq C\max(\|f\|_\infty, 1)\left(\frac{2^{2j\nu p}}{n^p} + \frac{2^{jp(2\nu+1)}}{n^{2p-1}}\right)^{1/p}. \tag{5.11}$$

By definition of j_1 and j_0 and since $\eta \leq 1/2$, one has that there exists a constant C such that for all $j_0 \leq j \leq j_1$, $\frac{2^j}{\sqrt{n}} \leq C\log(n)^\alpha$ which implies (since $p \geq 2$) that $\frac{2^{jp(2\nu+1)}}{n^{2p-1}} \leq C\frac{2^{2j\nu p}}{n^p}\log(n)^{\alpha p}$. By inserting this inequality in (5.11) one obtains that

$$\left(\mathbb{E}\left(\hat{\beta}_{jk}-\beta_{jk}\right)^{2p}\right)^{1/p} \leq C\max(\|f\|_\infty, 1)\max(\log(n)^\alpha, 1)\frac{2^{2j\nu}}{n}. \tag{5.12}$$

Now let $\gamma_{jk} = \sqrt{2\delta \log(n) \frac{\eta_j^2}{n^2} \hat{V}_{jk}} + \kappa \frac{\eta_j^2}{n^2} \delta \log(n)$, and remark that by definition of the threshold τ_{jk} and by using Lemmas 5.3 and 5.4 it follows that

$$\begin{aligned}
\mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \frac{\tau_{jk}}{\sqrt{1+\epsilon}}\right) &= \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \frac{\tau_{jk}}{\sqrt{1+\epsilon}}; \hat{V}_{jk} + \gamma_{jk} \geq V_{jk}\right) \\
&\quad + \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \frac{\tau_{jk}}{\sqrt{1+\epsilon}}; \hat{V}_{jk} + \gamma_{jk} \leq V_{jk}\right) \\
&\leq \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{\frac{2V_{jk}\delta \log(n)}{1+\epsilon}} + \frac{\eta_j}{3n}\frac{\delta \log(n)}{1+\epsilon}\right) \\
&\quad + \mathbb{P}\left(V_{jk} \geq \hat{V}_{jk} + \gamma_{jk}\right) \\
&\leq C(n^{-\frac{\delta}{1+\epsilon}} + n^{-\delta}) \leq C n^{-\frac{\delta}{1+\epsilon}} \qquad (5.13)
\end{aligned}$$

Now inserting (5.12) and (5.13) into inequality (5.10), and using the definition of j_1 and j_0 one finally obtains that for any $q > 1$ and $\epsilon > 0$

$$\begin{aligned}
\mathbb{E}Z &\leq C \max(\|f\|_\infty, 1) \max(\log(n)^\alpha, 1) \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} 2^{2j\nu} n^{-1-\frac{\delta}{q(1+\epsilon)}} \\
&\leq C \max(\|f\|_\infty, 1) \max(\log(n)^\alpha, 1) n^{-1-\frac{\delta}{q(1+\epsilon)}} 2^{j_1(2\nu+1)}.
\end{aligned}$$

By definition of j_1, $2^{j_1(2\nu+1)} \leq C n^{\eta \frac{2\nu+1}{\nu+1}} (\log n)^{\alpha(2\nu+1)} = C n^{\eta\left(1+\frac{\nu}{\nu+1}\right)} (\log n)^{\alpha(2\nu+1)}$ which implies that

$$\mathbb{E}Z \leq C \max(\|f\|_\infty, 1) \max(\log(n)^\alpha, 1) n^{-1+\eta\left(1+\frac{\nu}{\nu+1}\right)-\frac{\delta}{q(1+\epsilon)}} (\log n)^{\alpha(2\nu+1)}$$

By assumption, $\delta > \eta\left(1 + \frac{\nu}{\nu+1}\right)$. Hence there exists $\epsilon > 0$ such that $\delta > (1+\epsilon)\eta\left(1 + \frac{\nu}{\nu+1}\right)$, and one can then always find some $q > 1$ such that $\delta > q(1+\epsilon)\eta\left(1 + \frac{\nu}{\nu+1}\right)$. This implies that

$$\mathbb{E}Z \leq C \max(\|f\|_\infty, 1) \max(\log(n)^\alpha, 1) n^{-1} (\log n)^{\alpha(2\nu+1)}$$

which completes the proof. □

Now, using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ one has that

$$\mathbb{E}\tau_{jk}^2 \leq 4\delta \log(n) \mathbb{E}\tilde{V}_{jk} + 2\frac{\eta_j^2}{9n^2} \delta \log(n),$$

where $\tilde{V}_{jk} = \hat{V}_{j,k} + \sqrt{2\delta \log(n) \hat{V}_{j,k} \frac{\eta_j^2}{n^2}} + \delta \log(n) \kappa \frac{\eta_j^2}{n^2}$. Then using the inequality $2ab \leq a^2 + b^2$ and the fact that $\mathbb{E}\hat{V}_{j,k} = V_{jk}$, it follows that

$$\mathbb{E}\tilde{V}_{jk} \leq 3V_{jk} + (1+\kappa)\delta \log(n) \frac{\eta_j^2}{n^2},$$

which finally implies that there exists a constant C such that

$$\mathbb{E}\tau_{jk}^2 \leq C\left(\log(n) V_{jk} + (\log n)^2 \frac{\eta_j^2}{n^2}\right). \qquad (5.14)$$

Now, using Proposition 5.1, Lemma 5.5 and inequality (5.14), it follows that there exists two constants $C(\delta)$ and $C(\delta)'$ not depending on n and f, such that for any subset of indices $m \subset \Lambda_n$

$$\begin{aligned}\mathbb{E}\|\hat{f}_n - f\|^2 &\leq \sum_{k=0}^{2^{j_0}-1} \sigma_{j_0,k}^2 + C(\delta)\left(\sum_{(j,k)\in\Lambda_n\setminus m} \beta_{jk}^2 + \sum_{(j,k)\in m}(1+\log(n))V_{jk} + \sum_{j=j_1+1}^{+\infty}\sum_{k=0}^{2^j-1}\beta_{jk}^2\right) \\ &\quad + C'(\delta)\left(\max(\|f\|_\infty,1)\max(\log(n)^\alpha,1)\frac{(\log n)^{\alpha(2\nu+1)}}{n} + (\log n)^2\sum_{(j,k)\in m}\frac{\eta_j^2}{n^2}\right) \quad (5.15)\end{aligned}$$

and one can easily check that $\lim_{\delta\to\eta(1+\frac{\nu}{\nu+1})} C(\delta) = \lim_{\delta\to\eta(1+\frac{\nu}{\nu+1})} C'(\delta) = +\infty$ Then, as $\eta_j = \sum_{\ell\in C_j}|\tilde{\psi}_{jk}^\ell|$ it follows from (5.3) and (5.4) that $\eta_j^2 \leq C2^{j(2\nu+1)}$, which implies that for any $m \subset \Lambda_n$, $\sum_{(j,k)\in m}\frac{\eta_j^2}{n^2} \leq C\frac{1}{n^2}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}2^{j(2\nu+1)} \leq C\frac{2^{j_1(2\nu+2)}}{n^2} \leq Cn^{2(\eta-1)}(\log n)^{\alpha(2\nu+2)}$ by definition of j_1. Inserting this inequality in (5.15) with the model $m = \left\{(j,k); |\beta_{jk}|^2 \geq \log(n)V_{jk}\right\}$ finally yields that there exists two constants $C(\delta)$ and $C(\delta)'$ such that

$$\begin{aligned}\mathbb{E}\|\hat{f}_n - f\|^2 &\leq \sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2 + C(\delta)\left(\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{jk}^2,\log(n)V_{jk}) + \sum_{j=j_1+1}^{+\infty}\sum_{k=0}^{2^j-1}\beta_{jk}^2\right) \\ &\quad + C'(\delta)\left(\max(\|f\|_\infty,1)\max(\log(n)^\alpha,1)\frac{(\log n)^{\alpha(2\nu+1)}}{n} + \frac{(\log n)^{2+\alpha(2\nu+2)}}{n^{2(1-\eta)}}\right).\end{aligned}$$

To finally obtain inequality (5.15), remark that $V_{j,k} = \sigma_{j,k}^2 + \frac{1}{n}\beta_{j,k}^2$, and one can easily check that

$$\min(\beta_{jk}^2,\log(n)V_{jk}) \leq \min(\beta_{jk}^2,\log(n)\sigma_{jk}^2) + \frac{\log(n)}{n}\beta_{jk}^2,$$

which completes the proof of Theorem 3.2. □

5.5 Proof of Theorem 4.2

Given our assumptions, one has that $1 \leq p \leq 2$ which implies that $s^* = s + 1/2 - 1/p$. First we need the following lemma:

Lemma 5.6 *If $f \in B_{p,q}^s(A)$ with $1 \leq p \leq 2$ then*

$$\sum_{k=0}^{2^j-1}\beta_{jk}^2 \leq A^2 2^{-2js^*}. \quad (5.16)$$

Moreover, if $s > 1/p + 1/2$ with $1 \leq p \leq 2$, then there exists a constant $B > 0$ such that

$$\sup_{f\in B_{p,q}^s(A)}\|f\|_\infty \leq B.$$

Proof: since $f \in B_{p,q}^s(A)$ one has that $\sum_{k=0}^{2^{j_0}-1}|c_{j_0,k}|^p \leq A^p$ and $\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p \leq A^p 2^{-jp(s+1/2-1/p)}$. Since $p \leq 2$ it follows that

$$\left(\sum_{k=0}^{2^{j_0}-1}|c_{j_0,k}|^2\right)^{1/2} \leq \left(\sum_{k=0}^{2^{j_0}-1}|c_{j_0,k}|^p\right)^{1/p} \leq A, \quad (5.17)$$

and

$$\left(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^2\right)^{1/2} \leq \left(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p\right)^{1/p} \leq A2^{-j(s+1/2-1/p)}, \tag{5.18}$$

which proves the first part of the Lemma. Then, remark that

$$\|f\|_\infty \leq \|\sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}\|_\infty + \sum_{j=j_0}^{+\infty}\|\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}\|_\infty \tag{5.19}$$

Now, let $x \in [0,1]$ and remark that by Cauchy-Schwartz inequality

$$\left|\sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(x)\right|^2 \leq 2^{2j_0}\|\phi\|_\infty^2\left(\sum_{k=0}^{2^{j_0}-1}|c_{j_0,k}|^2\right),$$

and

$$\left|\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x)\right|^2 \leq 2^{2j}\|\psi\|_\infty^2\left(\sum_{k=0}^{2^{j_1}-1}|\beta_{j,k}|^2\right).$$

Hence the above inequalities, (5.17), (5.18), and (5.19) imply that

$$\|f\|_\infty \leq A2^{j_0}\|\phi\|_\infty + A\|\psi\|_\infty\sum_{j=j_0}^{+\infty}2^{-j(s-1/2-1/p)}$$

By assumption $s-1/2-1/p>0$ which implies that $\sum_{j=j_0}^{+\infty}2^{-j(s-1/2-1/p)} < +\infty$, and thus the result follows with $B = A\left(2^{j_0}\|\phi\|_\infty + \|\psi\|_\infty\sum_{j=j_0}^{+\infty}2^{-j(s-1/2-1/p)}\right)$. □

Note that although not always stated Lemma 5.6 will imply that the various bounds given below hold uniformly for $f \in D^s_{p,q}(A)$.

Let $R_n = R_1 + R_2 + R_3$ with

$$R_1 = \sum_{k=0}^{2^{j_0}-1}\sigma_{j_0,k}^2, \quad R_2 = \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1}\min(\beta_{jk}^2, \log(n)V_{jk}) \quad \text{and} \quad R_3 = \sum_{j=j_1+1}^{+\infty}\sum_{k=0}^{2^j-1}\beta_{jk}^2.$$

Given our assumptions on f and since $\eta = 1/2$, Theorem 3.2 and Lemma 5.6 imply that there exists two constants C and C' such that for every $n > \exp(1)$ and all $f \in D^s_{p,q}(A)$

$$\mathbb{E}\|\hat{f}_n - f\|^2 \leq CR_n + C'\frac{(\log n)^2}{n} \tag{5.20}$$

Hence to study the rate of convergence of $\hat{f}_n$, it suffices to study the asymptotic behavior of R_n. From Lemma 5.1 and by definition of j_0 one has that $R_1 \leq C\frac{2^{j_0(2\nu+1)}}{n} \leq C\frac{(\log n)^{(2\nu+1)}}{n}$ which implies that

$$R_1 = \mathcal{O}\left(n^{\frac{-2s}{2s+2\nu+1}}\right) \text{ (in the dense case) or } R_1 = \mathcal{O}\left(n^{\frac{-2s^*}{2s^*+2\nu}}\right) \text{ (in the sparse case)}. \tag{5.21}$$

Then remark that Lemma 5.6 implies that $R_3 = \mathcal{O}\left(2^{-2j_1 s^*}\right) = \mathcal{O}\left(n^{-\frac{2s^*}{2\nu+2}}(\log n)^{-2\alpha s^*}\right)$ by definition of j_1. Now remark that if $p=2$ then $s^* = s > 1$ (by assumption) and thus $\frac{2s^*}{2\nu+2} > \frac{2s}{2s+2\nu+1}$. If $1 \leq p < 2$ then

$s^* = s + 1/2 - 1/p$, and one can check that the condition $s > 1/2 + 1/p$ implies that $\frac{2s^*}{2\nu+2} > \frac{2s}{2s+2\nu+1}$ if $\nu(2-p) < ps^*$, and that $\frac{2s^*}{2\nu+2} > \frac{2s^*}{2s^*+2\nu}$ if $\nu(2-p) \geq ps^*$. Hence one obtains that

$$R_3 = \mathcal{O}\left(n^{-\frac{2s}{2s+2\nu+1}}\right) \text{ if } \nu(2-p) < ps^*, \tag{5.22}$$

$$R_3 = \mathcal{O}\left(n^{-\frac{2s^*}{2s^*+2\nu}}\right) \text{ if } \nu(2-p) \geq ps^*. \tag{5.23}$$

$$\tag{5.24}$$

Now it remains to study the term R_2.

For this let us consider first the dense case when $\nu(2-p) < ps^*$, and decompose $R_2 = R_{21} + R_{22}$ with

$$R_{21} = \sum_{j=j_0}^{j_2} \sum_{k=0}^{2^j-1} \min(\beta_{jk}^2, \log(n) V_{jk}) \text{ and } R_{22} = \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} \min(\beta_{jk}^2, \log(n) V_{jk}),$$

where $j_2 = j_2(n)$ is the integer such that $2^{j_2} > (n/\log(n))^{\frac{1}{2s+2\nu+1}} \geq 2^{j_2-1}$ (note that given our assumptions $j_2 \leq j_1$ for all sufficiently large n). Then remark that by Lemma 5.1, $R_{21} \leq \sum_{j=j_0}^{j_2} \sum_{k=0}^{2^j-1} \log(n) V_{jk} \leq C \sum_{j=j_0}^{j_2} \log(n) \frac{2^{j(2\nu+1)}}{n} \leq C \log(n) \frac{2^{j_2(2\nu+1)}}{n} \leq C(n/\log(n))^{-\frac{2s}{2s+2\nu+1}}$. Hence

$$R_{21} = \mathcal{O}\left((n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right) \tag{5.25}$$

Now remark that if $p \geq 2$ then by equation (5.16) and by definition of j_1 and j_2 it follows that $R_{22} \leq C \sum_{j=j_2+1}^{j_1} 2^{-2js} \leq C\left(n^{-\frac{2s}{2\nu+2}} - (n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right)$ which implies that

$$R_{22} = \mathcal{O}\left((n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right). \tag{5.26}$$

Now if $1 \leq p < 2$ remark that R_{22} can be written as

$$\begin{aligned} R_{22} &= \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \mathbb{1}_{\{\beta_{jk}^2 < \log(n) V_{jk}\}} + \log(n) V_{jk} \mathbb{1}_{\{\beta_{jk}^2 > \log(n) V_{jk}\}} \\ &= \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} |\beta_{jk}|^{2-p} |\beta_{jk}|^p \mathbb{1}_{\{\beta_{jk}^2 < \log(n) V_{jk}\}} + (\log(n) V_{jk})^{1-p/2} (\log(n) V_{jk})^{p/2} \mathbb{1}_{\{\beta_{jk}^2 > \log(n) V_{jk}\}} \\ &\leq 2 \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} (\log(n) V_{jk})^{1-p/2} |\beta_{jk}|^p. \end{aligned}$$

By Lemma 5.1 $V_{jk} \leq C \frac{2^{2j\nu}}{n}$, and since $f \in B_{p,q}^s(A)$ if follows that there exists a constant C depending only on p, q, s, A such that $\sum_{k=0}^{2^j-1} |\beta_{jk}|^p \leq C 2^{-js^*p}$, which implies that

$$R_{22} \leq C(n/\log(n))^{-1+p/2} \sum_{j=j_2+1}^{j_1} 2^{j(2\nu - \nu p - s^* p)}.$$

Now in the dense case, one has that $2\nu - \nu p - s^* p < 0$ which implies that

$$R_{22} = \mathcal{O}\left((n/\log(n))^{-1+p/2} 2^{j_2(2\nu - \nu p - s^* p)}\right) = \mathcal{O}\left((n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right) \tag{5.27}$$

by using the definition of j_2. Hence combining (5.25), (5.26) and (5.27) it follows that in the dense case for $1 \leq p \leq \infty$

$$R_2 = \mathcal{O}\left((n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right). \tag{5.28}$$

Now consider the sparse case when $\nu(2-p) \geq ps^*$, and decompose $R_2 = R_{21} + R_{22}$ with

$$R_{21} = \sum_{j=j_0}^{j_2} \sum_{k=0}^{2^j-1} \min(\beta_{jk}^2, V_{jk}) \text{ and } R_{22} = \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} \min(\beta_{jk}^2, V_{jk}),$$

where $j_2 = j_2(n)$ is the integer such that $2^{j_2} > (n/\log(n))^{\frac{1}{2s^*+2\nu}} \geq 2^{j_2-1}$. Note that in the sparse case then necessarily $1 \leq p < 2$, and as previously one can thus remark that R_{21} can be written as

$$R_{21} = \sum_{j=j_0}^{j_2} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \mathbb{1}_{\{\beta_{jk}^2 < \log(n) V_{jk}\}} + \log(n) V_{jk} \mathbb{1}_{\{\beta_{jk}^2 > \log(n) V_{jk}\}} \leq 2 \sum_{j=j_0}^{j_2} \sum_{k=0}^{2^j-1} (\log(n) V_{jk})^{1-p/2} |\beta_{jk}|^p.$$

Now using again the fact that $V_{jk} \leq C\frac{2^{2j\nu}}{n}$ (by Lemma 5.1), that $\sum_{k=0}^{2^j-1} |\beta_{jk}|^p \leq C 2^{-js^*p}$ (since $f \in B_{p,q}^s(A)$), and by definition of j_2 and j_0 it follows that

$$R_{21} \leq C(n/\log(n))^{-1+p/2} \left(2^{j_2(2\nu - p\nu - ps^*)} - 2^{j_0(2\nu - p\nu - ps^*)}\right) = \mathcal{O}\left((n/\log(n))^{-\frac{2s^*}{2s^*+2\nu}}\right). \tag{5.29}$$

Then, remark that by equation (5.16), $R_{22} \leq \sum_{j=j_2+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \leq \sum_{j=j_2+1}^{j_1} 2^{-2js^*}$. Hence, $R_{22} = \mathcal{O}\left(2^{-2j_2 s^*}\right)$, and thus by definition of j_2

$$R_{22} = \mathcal{O}\left((n/\log(n))^{-\frac{2s^*}{2s^*+2\nu}}\right). \tag{5.30}$$

Finally combining (5.29) and (5.30) it follows that in the dense case

$$R_2 = \mathcal{O}\left((n/\log(n))^{-\frac{2s^*}{2s^*+2\nu}}\right). \tag{5.31}$$

Then, combining (5.21), (5.22), (5.23), (5.28) and (5.31) implies that $R_n = \left((n/\log(n))^{-\frac{2s}{2s+2\nu+1}}\right)$ in the dense case, and $R_n = \left((n/\log(n))^{-\frac{2s^*}{2s^*+2\nu}}\right)$ in the sparse case. Using inequality (5.20) this completes the proof of Theorem 4.2. □

References

Antoniadis, A., Grégoire, G. and Nason, G. (1999). Density and hazard rate estimation for right censored data using wavelet methods. *J. R. Statist. Soc. Series B*, *61*, 63-84.

Bigot, J. and Van Bellegem, S. (2009). Log-density deconvolution by wavelet thresholding. *Scandinavian Journal of Statistics*, *to appear*.

Buckheit, J., Chen, S., Donoho, D. and Johnstone, I. (1995). *Wavelab reference manual* (Tech. Rep.). Department of Statistics, Stanford University. (`http://www-stat.stanford.edu/software/wavelab`)

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparse density estimation with l1 penalties. *Lecture Notes in Artificial Intelligence (COLT 2007), Springer*, 530 - 544.

Candès, E. (2005). Modern statistical estimation via oracle inequalities. *Acta Numerica*, *15*, 257-325.

Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, *83*, 1184–1186.

Castellan, G. (2003). Density estimation via exponential model selection. *IEEE Transactions on Information Theory*, *49*, 2052-2060.

Cavalier, L., Golubev, G. K., Picard, D. and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, *30*(3), 843–874. (Dedicated to the memory of Lucien Le Cam)

Comte, F., Rozenholc, Y. and Taupin, M.-L. (2006a). Finite sample penalization in adaptive density deconvolution. *J. Stat. Comput. Simul.*, to appear.

Comte, F., Rozenholc, Y. and Taupin, M.-L. (2006b). Penalized contrast estimator for density deconvolution. *Canad. J. Statist.*, *34*, XXX.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, *81*, 425Ð455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B*, *57*, 301–369.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, *24*, 508–539.

Efromovich, S. (2008). Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.*, *3*, 1127-1155.

Fan, J. (1991). On the optimal rate of convergence for nonparametric deconvolution problems. *Ann. Statist.*, *19*, 1257–1272.

Fan, J. and Koo, J.-Y. (2002). Wavelet deconvolution. *IEEE Trans. Inform. Theory*, *48*, 734–747.

Herrick, D., Nason, G. and Silverman, B. (2001). Some new methods for wavelet density estimation. *Sankhya, Series A*, *63*, 394-41.

Johannes, J. (2008). Deconvolution with unknown error distribution. *Ann. Statist., to appear*.

Johnstone, I., Kerkyacharian, G., Picard, D. and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *J. Roy. Statist. Soc. Ser. B*, *66*, 547–573.

Johnstone, I. M. (2002). Function estimation in gaussian noise: Sequence models. *Unpublished Monograph, http://www-stat.stanford.edu/imj/*.

Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli*, *10*, 187-220.

Kolaczyk, E. (1994). *Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data.* Ph.d. thesis, Department of Statistics, Stanford University, Stanford.

Koo, J. and Kooperberg, C. (2000). Logspline density estimation for binned data. *Statistics and Probability Letters*, *46*, 133-47.

Koo, J.-Y. (1999). Logspline deconvolution in Besov space. *Scand. J. Statist.*, *26*, 73–86.

Kosarev, E., Shul'man, A., Tarasov, M. and Lindstroem, T. (2003). Deconvolution problems and superresolution in Hilbert-transform spectroscopy based on a.c. Josephson effect. *Comput. Phys. Comm.*, *151*, 171–186.

Masry, E. (2003). Deconvolving multivariate kernel density estimated from contaminated associated observations. *IEEE Trans. Inform. Theory*, *49*, 2941–2952.

Massart, P. (2006). *Concentration inequalities and model selection: Ecole d'été de probabilités de saint-flour xxxiii - 2003.* Lecture Notes in Mathematics Springer.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge: Cambridge University Press.

Pensky, M. and Sapatinas, T. (2008). Functional deconvolution in a periodic setting: uniform case. *Ann. Statist.*, *to appear*.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, *27*, 2033–2053.

Postel-Vinay, F. and Robin, J.-M. (2002). Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica*, *70*, 2295–2350.

Raimondo, M. and Stewart, M. (2007). The waved transform in r: performs fast translation-invariant wavelet deconvolution. *Journal of Statistical Software*, *21*(3), 1-27.

Reynaud-Bouret, P. and Rivoirard, V. (2008). Adaptive thresholding estimation of a poisson intensity with infinite support. *preprint*.

Rigollet, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli*, *12*(2), 351–370.

Rosenthal, H. P. (1972). On the span in L^p of sequences of independent random variables. II. 149–167.

Vidakovic, B. (1999). *Statistical modeling by wavelets.* New York: Wiley.